

Strength of Evidence

Grinnell College

March 30, 2026

Warm-up

The culmen (upper bill) of 20 male sharp-shinned hawks were sampled and measured, generating the following statistics:

$$\bar{x} = 10.81, \quad \hat{\sigma} = 3.93$$

With this information, respond to the following:

1. Construct a 90% confidence interval for the average culmen length of male sharp-shinned hawks
2. Construct a t -statistic for the culmen under the null hypothesis $H_0 : \mu = 12.5$
3. Is the value of this hypothesis contained within your 90% CI?
4. Compare your t -statistic to the critical value found in (1). Based on this, what conclusion would you come to with your hypothesis?

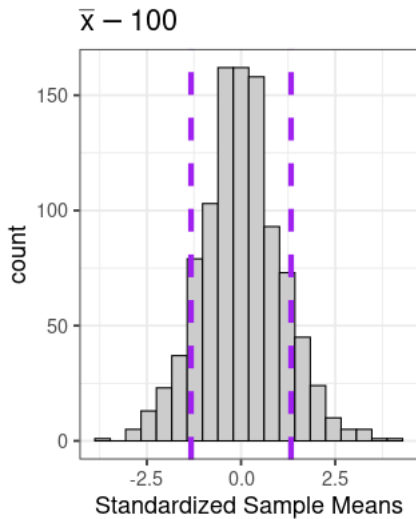
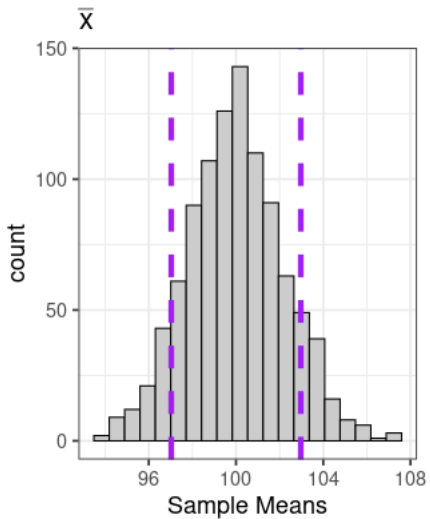
Last week we introduced the idea of the **null distribution**

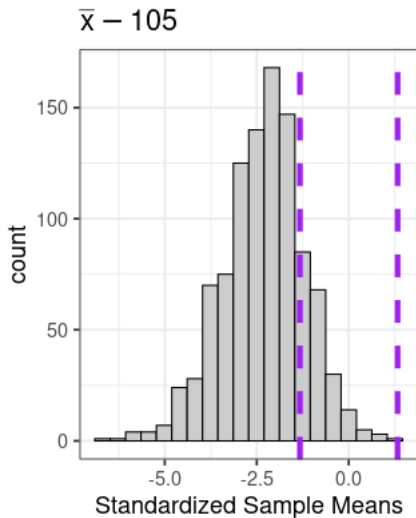
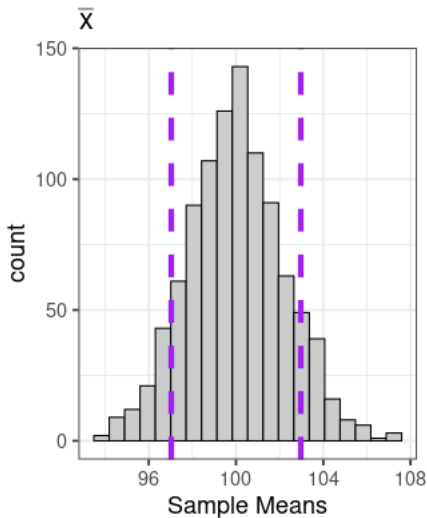
If we were to collect many samples of \bar{X} , the null distribution refers to the distribution of statistics

$$t = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

when $H_0 : \mu = \mu_0$ is true.

When H_0 is not true, the distribution of our t statistics will *not* follow a t distribution as it will no longer be centered at zero





Consider the t

Consider the pieces of a t-statistic

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

1. $\bar{x} - \mu_0$ indicates the distance between my observed data and my null hypothesis, though we cannot use this alone as it does not include a degree of “certainty” associated with \bar{x}
2. $\hat{\sigma}$ is my estimate of the population’s standard deviation. When this is large, there will be more uncertainty in my estimate of \bar{x}
3. n represents the number of observations in my sample – the more observations I have, the more confidence I will have in my estimate

We should have a sense of how each of these components impact my t -statistic

Consider the t

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

We can think, then, of the t -statistic as being a measure of evidence *against* the null hypothesis

If:

1. \bar{x} is far from μ and
2. Our certainty in \bar{x} is high (e.g., low $\hat{\sigma}$ or large n)

then our statistic t will be larger. A larger t statistic is less likely than a smaller one

What we consider “large” will depend on the t -distribution

t-distribution

When the null hypothesis is true,

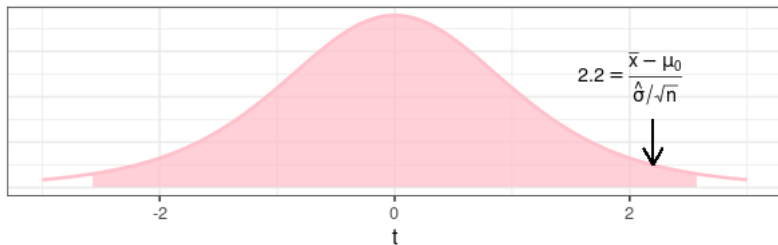
$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

follows a t -distribution with $n - 1$ degrees of freedom

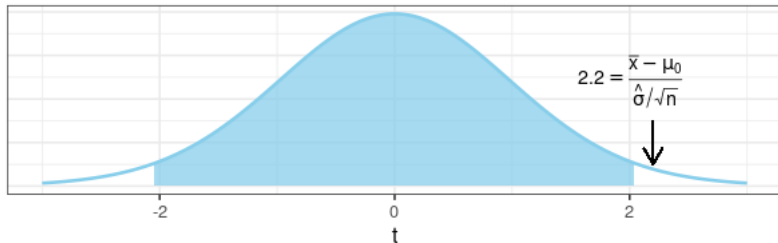
The degrees of freedom tells us, relatively speaking, what values are considered “large”

$t = 2.2$ may be considered “large” when $df = 30$ but not when $df = 5$

95% of a t-distribution with df = 5



95% of a t-distribution with df = 30



In summary

The process (so far) goes like this:

1. Assume our null hypothesis $H_0 : \mu = \mu_0$ is true
2. Compute a t -statistic with our observed data
3. Ask: is this t -statistic “large”?
 - ▶ This will depend on the degrees of freedom
 - ▶ It will also depend on our confidence level and critical value
4. Either reject or fail to reject depending on how our t statistics compares to the relevant critical value

What we would like now is a way to *quantify* the strength of our evidence without comparing our statistic to a particular critical value

p-values

Where *critical values* allow us to move from percentiles to a value, a **p-value** takes us from a value to a percentile

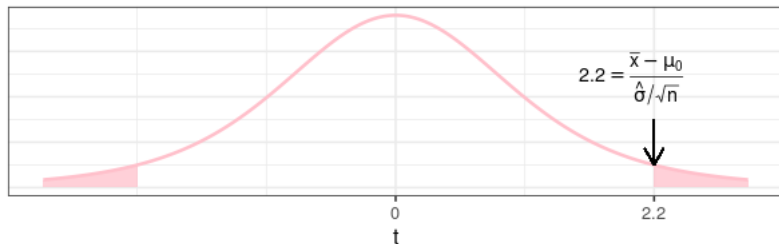
More specifically, a p-values asks: “If the null hypothesis ($H_0 : \mu = \mu_0$) is true, what is the probability that we have observed our data *or something at least as large*”

The proportion of our null distribution at least as large as our observed data is expressed as a probability which we call the p-value:

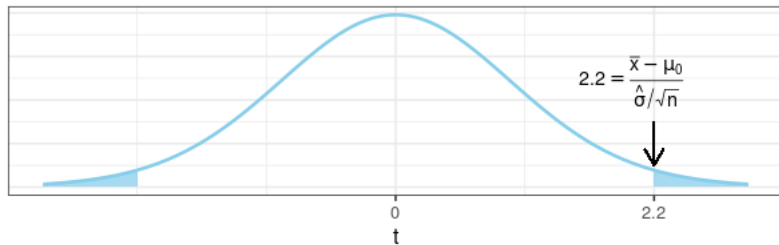
$$\text{p-value} = P(\text{observed data} \mid H_0 \text{ is true})$$

pvalue stuff

p-value with df = 5

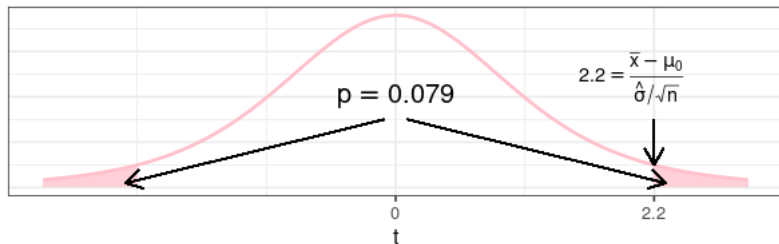


p-value with df = 30

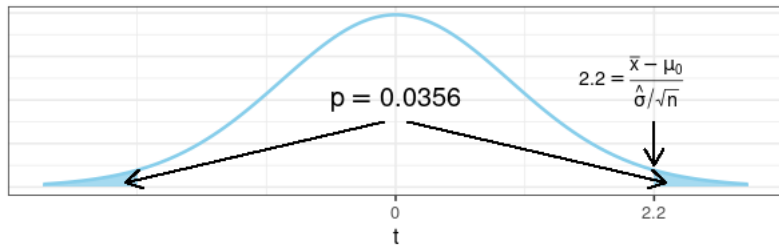


pvalue stuff

p-value with df = 5



p-value with df = 30



pvalue

Just as our t -statistic is based on the null hypothesis, so, too, is our p -value

The sampling distribution under the null hypothesis is called our **null distribution**

The p -value, then, is an indication of how likely (or unlikely) our observed data is *under the assumption that the null hypothesis is true*

If we were to have a different null hypothesis, our p -value would be different as well

P-values and decision making

Last week we saw that making a decision by comparing our t -statistic with, say, a 90% critical value was associated with a 10% **error rate**

Likewise, comparing our t -statistic with a 95% confidence interval provides us with a 5% error rate

This error rate, called the **Type I Error Rate**, expresses the probability of incorrectly rejecting the null hypothesis when H_0 is true

We represent this error rate with the Greek letter α . So, for a 95% confidence interval, we are conducting a hypothesis test with $\alpha = 0.05$

p-values and decision making

For a given error rate α , there is an equivalence between checking p-values and comparing test statistics with the associated critical values:

1.

$$p < \alpha \Leftrightarrow C < |t|$$

2.

$$p > \alpha \Leftrightarrow |t| < C$$

Whereas checking $|t| < C$ gives us a binary decision, assessing the p -value allows us to quantify the strength of our evidence

Example

Suppose we are interested in testing the hypothesis that the true average hallux length for male sharp-shinned hawks is 11mm. To this end, we collected two samples:

- ▶ **Sample 1:** $\bar{x}_1 = 15.61$, $\hat{\sigma} = 6.72$, $n = 10$
- ▶ **Sample 2:** $\bar{x}_2 = 13.61$, $\hat{\sigma} = 6.12$, $n = 25$

We might notice in passing that:

1. Sample 1 has an observed sample mean that is further away from μ_0
2. Sample 2 has more than double the observations as Sample 1
3. The observed variability in both samples is about the same

t-statistics and p-values

We can start by constructing t -statistics for each of our samples

Sample 1:

$$t = \frac{15.61 - 11}{6.72/\sqrt{10}} = 2.17$$

Sample 2:

$$t = \frac{13.61 - 11}{6.12/\sqrt{25}} = 2.13$$

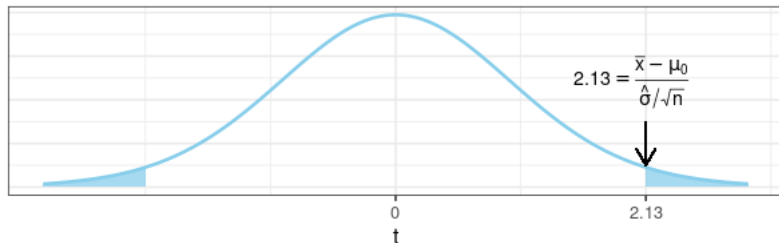
We might conclude that, having the larger t -statistic that Sample 1 provides more evidence against the null; however, the *null distribution* for each statistic is different according to its degrees of freedom

p-values

p-value with df = 9

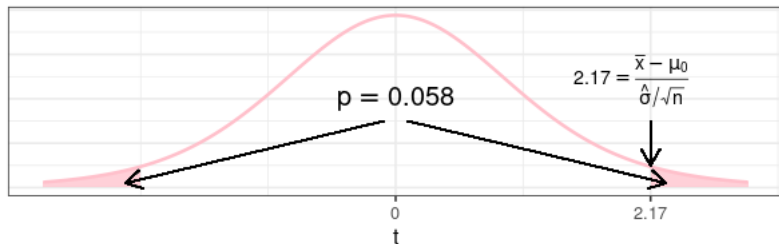


p-value with df = 24

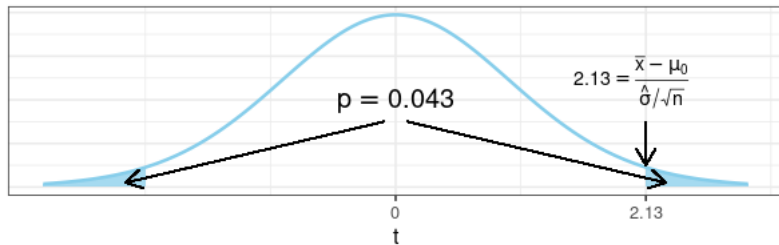


p-values

p-value with df = 15



p-value with df = 40



Drawing Conclusions

Suppose we were wishing to test our hypothesis $H_0 : \mu = 11$ with each of our two samples with a Type I error rate of $\alpha = 0.05$

Sample 1:

With a t statistic of $t = 2.17$ following a null distribution with $df = 9$, we find a p -value of $p = 0.058$. Since $p > \alpha$, we fail to reject our null hypothesis

Sample 2:

With a t statistic of $t = 2.13$ following a null distribution with $df = 24$, we find a p -value of $p = 0.043$. Since $p < \alpha$, we reject our null hypothesis

Relationship Between Critical Values and α

As we noted before, we can equivalently check the test statistics for each sample against the appropriate critical values for each null distribution's degrees of freedom:

$$C_9 = 2.262 \quad C_{24} = 2.063$$

Immediately, we see that for our first sample, $t = 2.17 < C_9$, telling us that our observed data is within the middle 95% and we would fail to reject

Likewise for our second sample, $t = 2.13 > C_{24}$, indicating that we *would reject*

Relationship between CI and α

Finally consider the confidence intervals themselves:

Sample 1:

$$15.61 \pm 2.262 \times \left(6.72/\sqrt{10}\right) = (10.8, 20.4)$$

Sample 2:

$$13.61 \pm 2.063 \times \left(6.12/\sqrt{25}\right) = (11.1, 16.2)$$

Here, we see that the null hypothesis $H_0 : \mu = 11$ is contained within the 95% confidence interval of Sample 1, indicating that we fail to reject, while it is *not* within the interval for Sample 2, indicating rejection

A Note

p -values are useful in that they allow us to compare the size of different t -statistics, relative to the null distribution

In the special case where two samples have the same sample size, we can simply compare the t -statistics directly. Why?

p-value facts

p -values are notorious for how easily they are misinterpreted. Here are a few key facts:

- ▶ A p -value *is not* the probability that the null hypothesis is false
- ▶ A p -value *is not* the probability of having observed our data by random chance
- ▶ A p -value *does not* tell us the magnitude of difference ($\bar{x} - \mu_0$) or the size of the effect
- ▶ A p -value *must* be taken in the context of the study; a p -value of 0.05 is completely arbitrary
- ▶ A p -value *is* a probabilistic statement relating observed data to a hypothesis

Hypothesis testing involves formulating statements about our population and then checking the consistency of our hypothesis with observed data

Rather than getting a binary yes/no answer by checking t -statistics with a specific critical value, a **p-value** allows us to *quantify* to what extent our observed data is consistent with a null hypothesis

There is a one-to-one relationship between critical values and our Type I error rate, α

Checking that $t < C$ is equivalent to checking if $p < \alpha$