Strength of Evidence Revisited

Grinnell College

April 9, 2025

Warm-up

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 three samples:

- **Sample 1:** $\overline{x}_1 = 12.3178$, $\hat{\sigma} = 5$, $n = 25$
- **Sample 2:** $\overline{x}_2 = 12.7109$, $\hat{\sigma} = 5$, $n = 25$
- **Sample 2:** $\overline{x}_3 = 12.5109$, $\hat{\sigma} = 5$, $n = 25$

Provide the following additional information:

1. What are the critical values associated with 80% and 90% confidence?
2. What are the error rates associated with 80% and 90% confidence
3. Draw a t-distribution on a sheet of paper and mark where each of these critical values lay (leave plenty of room)

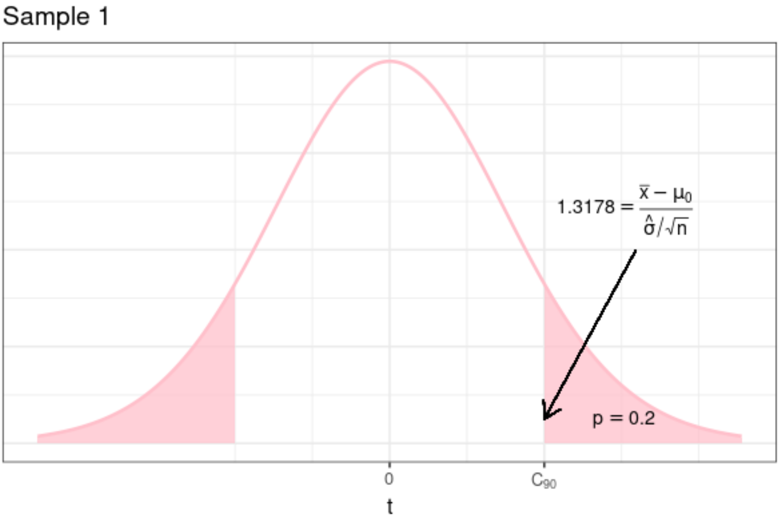
Sample 1

From Sample 1, we find $\overline{x}_1 = 12.3178$, $\hat{\sigma} = 5$, $n = 25$. The associated t-statistic is

$$t = \frac{12.3178 - 11}{5/\sqrt{25}} = 1.3178$$

- What decision do we make regarding H_0 at 80% confidence?
- What decision do we make regarding H_0 at 90% confidence?
- What is the greatest amount of confidence we could obtain while still rejecting H_0?
- Where does this t-statistic fall relative to the CV you drew in the warm-up?

Sample 1



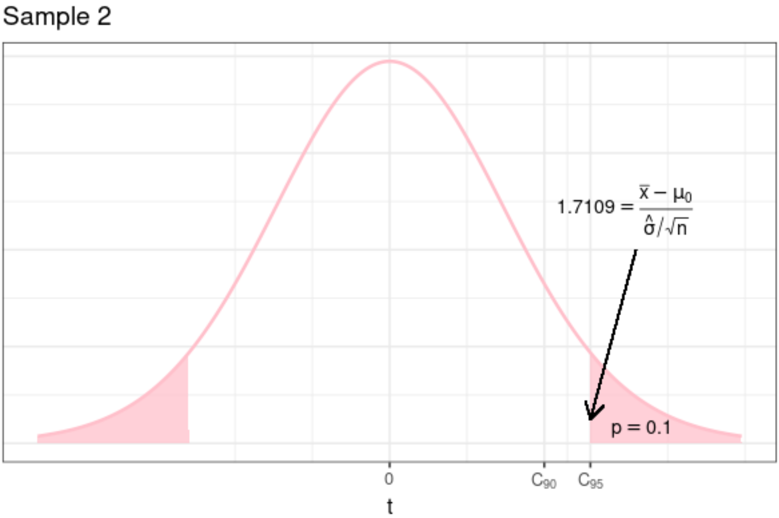
Sample 2

From Sample 1, we find $\overline{x}_2 = 12.7109$, $\hat{\sigma} = 5$, $n = 25$. The associated t-statistic is

$$t = \frac{12.7109 - 11}{5/\sqrt{25}} = 1.7109$$

- What decision do we make regarding H_0 at 80% confidence?
- What decision do we make regarding H_0 at 90% confidence?
- What is the greatest amount of confidence we could obtain while still rejecting H_0?
- Where does this t-statistic fall relative to the CV you drew in the warm-up?

Sample 2



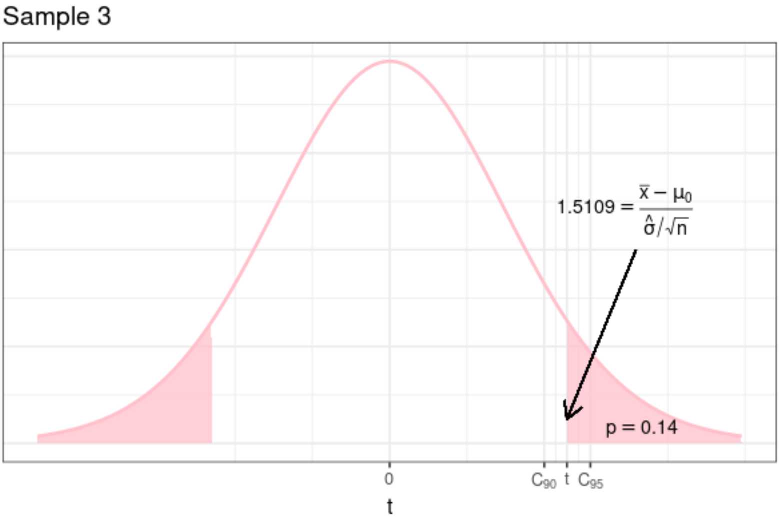
Sample 3

From Sample 1, we find $\overline{x}_3 = 12.5109$, $\hat{\sigma} = 5$, $n = 25$. The associated t-statistic is

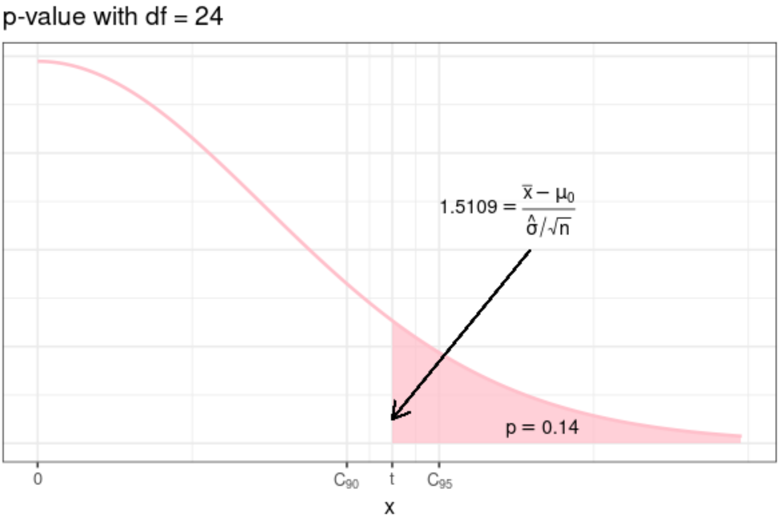
$$t = \frac{12.7109 - 11}{5/\sqrt{25}} = 1.5109$$

- What decision do we make regarding H_0 at 80% confidence?
- What decision do we make regarding H_0 at 90% confidence?
- What is the greatest amount of confidence we could obtain while still rejecting H_0?
- Where does this t-statistic fall relative to the CV you drew in the warm-up?

Sample 3



Sample 3



p-value with df = 24
$1.5109 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
p = 0.14
0
C_{90}
t
C_{95}
x

t-distribution

When the null hypothesis is true,

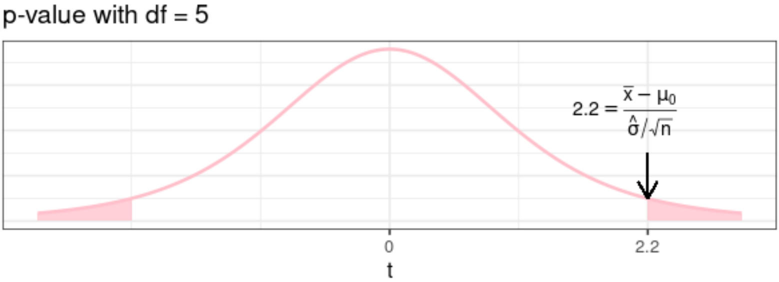
$$t = \frac{\overline{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

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

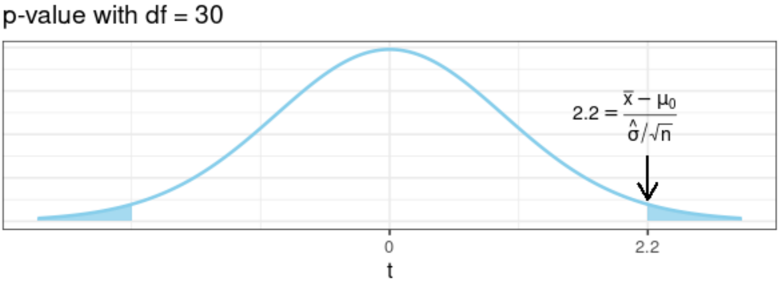
The degrees of freedom tells us, relatively speaking, what values are considered "large"

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

Sample 3

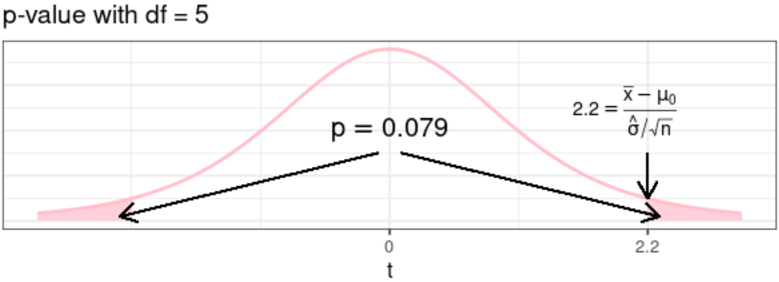


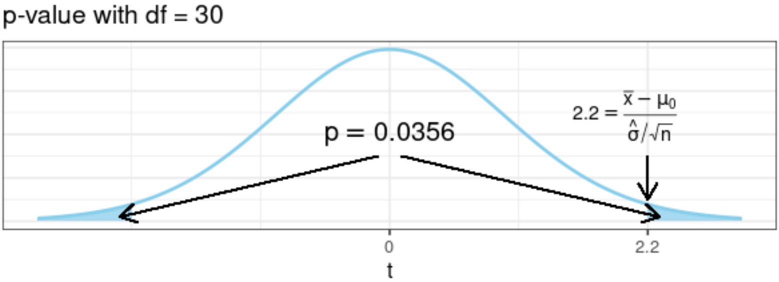
p-value with df = 5
$2.2 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
0
2.2
t



p-value with df = 30
$2.2 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
0
2.2
t

Sample 3


p-value with df = 5
p = 0.079
$2.2 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
0
2.2
t


p-value with df = 30
p = 0.0356
$2.2 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
0
2.2
t

Why p-values

- Quantifies t-statistic without reference to critical values or degrees of freedom
- Instead allows me to specify an error rate and come to conclusions accordingly

Make note that each value of C corresponds to an error rate, α, which is equal to 1 minus the confidence i.e., a 90% confidence corresponds to a 10% error rate

Likewise, each t statistic corresponds to a p-value. When $t > C$, it will follow that $p < \alpha$

Example (Revisited

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

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:

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

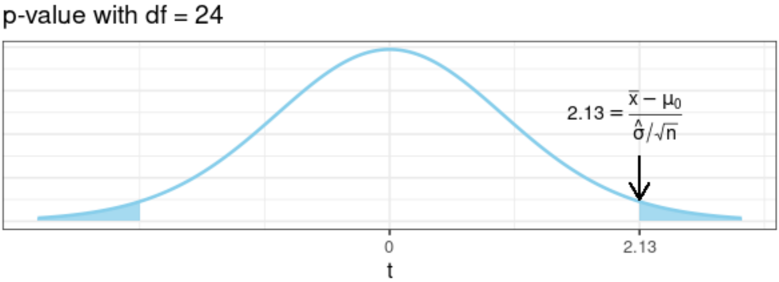
Sample 2:

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

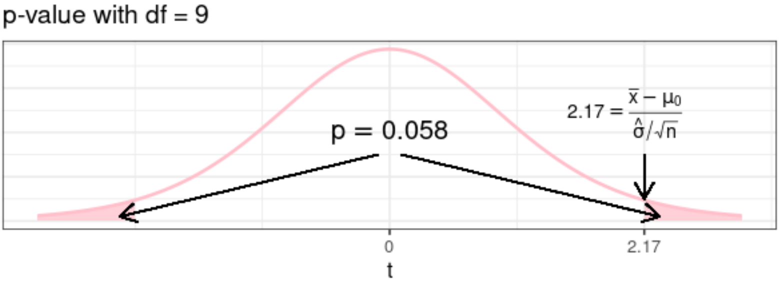
We might conclude that, having the larger t-statistic that Sample 1 provides more evidence against the null; however, the *null distribution* for each statistic is different according to its degrees of freedom

p-values

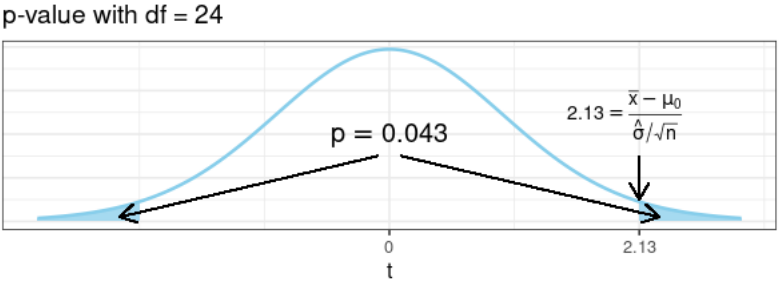

p-value with df = 9
$2.17 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
0
2.17
t


p-value with df = 24
$2.13 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
0
2.13
t

p-values



p-value with df = 9
p = 0.058
$2.17 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
0
2.17
t



p-value with df = 24
p = 0.043
$2.13 = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
0
2.13
t

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 CI and α

If we were to construct 95% confidence intervals with our observed data, we would first find critical values for each of our null distributions, according to sample size (note: C_{df} refers to critical value with *df* degrees of freedom):

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

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

Now 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

Review

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 $p < \alpha$