Math3350: Two-sample t procedures for confidence intervals

Jesse Wheeler

2025-04-14

Logistics

- Any questions about previous material?

Motivation

Comparing samples from two populations

We often have data (or samples) collected on two distinct populations or treatments. Our goal is to use samples to compare the sames from these independent groups.

Example 1:

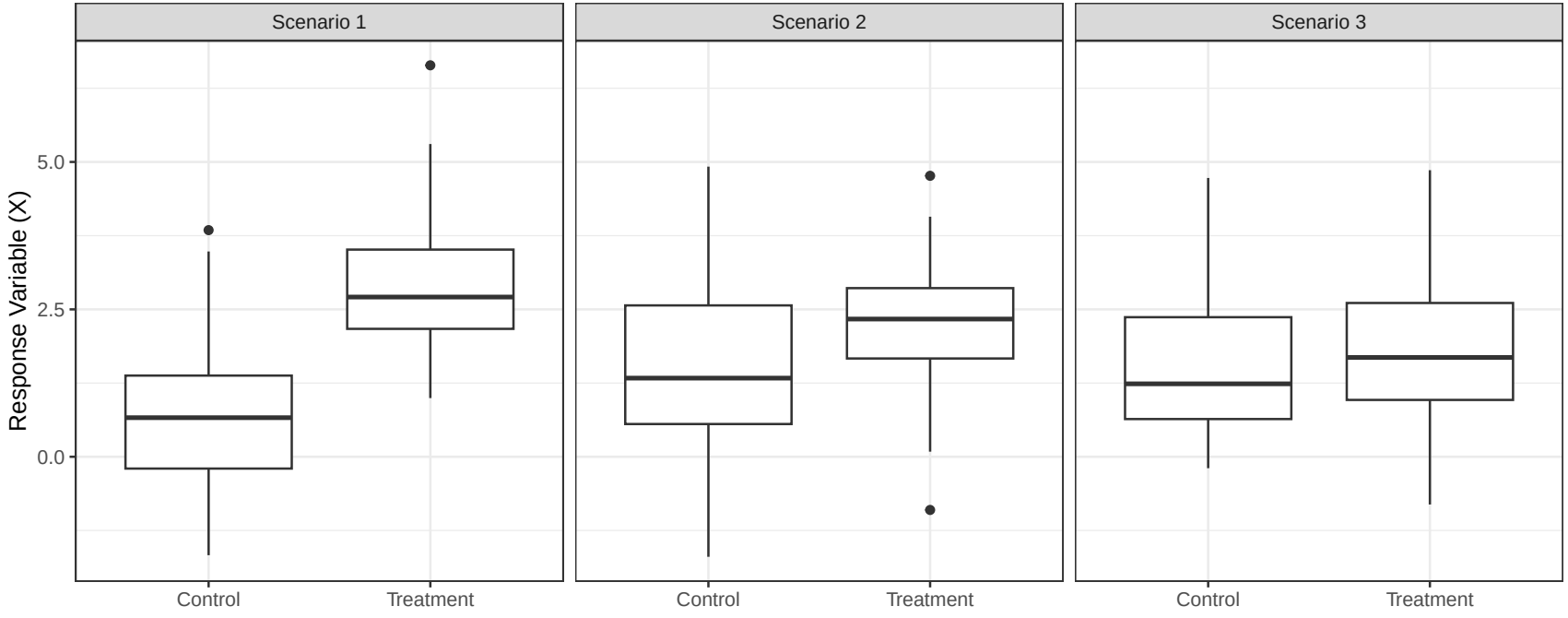
- Researchers used an animal model to examine the long-term impact of exposure to *tricolosan*, a broad-spectrum antimicrobial agent commonly added to soaps. They randomly assigned mice to diets containing either 0.08% triclosan or no triclosan for eight months, then compared the liver weights in each group.
  - The response variable for each group X_1 (treatment), X_0 (control) is liver weight.

Example 2:

- Researchers selected a random sample of 682 medical records of California children in the same age group who had received care from the same clinicians on the same day but were not diagnosed with pertussis. The researchers then compared the length of time since vaccination in both groups.
  - The response variable for each group X_i is the length of time since vaccination, the factor level i corresponds to pertussis diagnosis.

Plotting the response variable

One way to compare groups is plot them. For instance, we could consider doing histograms, boxplots, or dotplots. Below is an example of comparing samples using boxplots.



Scenario 1

The samples are different enough that we can clearly see a difference between the groups.

Scenario 2

We can still see a difference, but the difference not as obvious.

Scenario 3

The plots are not the exact same, but we only have a finite sample. Are we sure the populations are really that different?

Plotting

Plots are useful to help compare the shares, centers, and spreads of the two samples.

Question: What is the problem with making scientific conclusions using figures like those above?

Two-sample t procedures.

Setup

- Ideally, we would have a theoretically supported, objective method for comparing group means. A class of approaches that satisfy these ideals in practice is "Two-sample t procedures".
- In what follows, we suppose there are two groups of interest, Group 1 and Group 2, and that we have obtained samples of some response variable for each of these groups.

Setup

- We denote $\bar{X}_1$ and $\bar{X}_2$ be the sample mean of the response variables for the respective groups,
- Similarly, s_1 and s_2 are the sample standard deviations.

Population	Mean	SD	Sample Mean	Sample SD
1	μ_1	σ_1	$\bar{X}_1$	s_1
2	μ_2	σ_2	$\bar{X}_2$	s_2

Unknowns

There are four unknown parameters: population means (μ_1, μ_2), and standard deviations (σ_1, σ_2).

Our goal is to compare the two population means in a mathematically rigorous way. Specifically, we would like to test the Hypothesis:

$$H_0 : \ \mu_1 = \mu_2,$$

or equivalently,

$$H_0 : \ \mu_1 - \mu_2 = 0$$

against the alternative

$$H_1 : \ \mu_1 - \mu_2 \neq 0,$$

or

$$H_1 : \ \mu_1 - \mu_2 > 0 \qquad \text{or} \qquad H_1 : \ \mu_1 - \mu_2 < 0$$

Confidence Intervals

There is a duality between hypothesis tests and confidence intervals, meaning that the same sampling distributions used for a hypothesis test can be leveraged for creation of confidence intervals.

Today, we will use this to get a confidence interval for the population value $\mu_1 - \mu_2$.

Conditions

T be able to use a two-sample t-procedure, we need the following conditions:

- We have two simple random samples (SRS). This means we need *independence* between observations, or that observations from one sample are unrelated to the observations in the other sample.
- Both samples come from *approximately normal* distributions. In practice, it is often enough that the distributions have similar shapes, are roughly symmetric, and have no strong outliers.

Sampling Distributions

Recall that $\bar{X}_1$ and $\bar{X}_2$ are only *sample statistics*. Therefore, the difference we actually compute

$$\bar{X}_1 - \bar{X}_2$$

is specific to the independent samples we obtained. A different set of random samples would likely yield a different value for $\bar{X}_1 - \bar{X}_2$. How much the difference can vary from sample to sample is given by its *sampling distribution*.

While the proof is beyond the scope of this course, it can be shown that the distribution of $\bar{X}_1 - \bar{X}_2$ has a mean $\mu_1 - \mu_2$, and a standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Question: What can notice about the standard deviation as the number of samples of both factors increases?

t-statistic

When both populations are Normally distributed, the sample distributions of $\bar{X}_1$ and $\bar{X}_2$ will also be Normally distributed, and so will their difference $\bar{X}_1 - \bar{X}_2$:

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

μ_i and σ_1 are unknown. However, we are interested in estimating the difference $\mu_1 - \mu_2$, with our null hypothesis being $\mu_1 - \mu_2 = 0$.

Approximating standard deviation

In practice, don't know σ_1 and σ_2, so we estimate them by the sample standard deviations, s_1, s_2. Using these substitutions, the estimated standard deviation is known as the *standard error*:

$$\text{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The sample statistic that we are interested in is the standardized version of the sample difference:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\text{SE}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

We call this the *two-sample t statistic*, and it has approximately a t distribution.

Degrees of freedom

The degrees of freedom (df) is given by:

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

This is usually a good approximation if the populations are approximately normally distributed and if n_1 and n_2 are 5 or larger.

Unlike in the one-sample t-test, the df here depends on both the number of observations (n_1, n_2), and the sample variance (s_1^2, s_2^2). Notably, it's usually not a whole number.

The exact value is **always** between $\min\{n_1 - 1, n_2 - 1\}$ and $n_1 + n_2 - 2$.

Confidence Intervals

Like the one-sample situation, we can use the properties of the t-distribution to obtain a confidence interval about the value $\mu_1 - \mu_2$. Specifically, a level-C confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{X}_1 - \bar{X}_2) \pm t_C \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

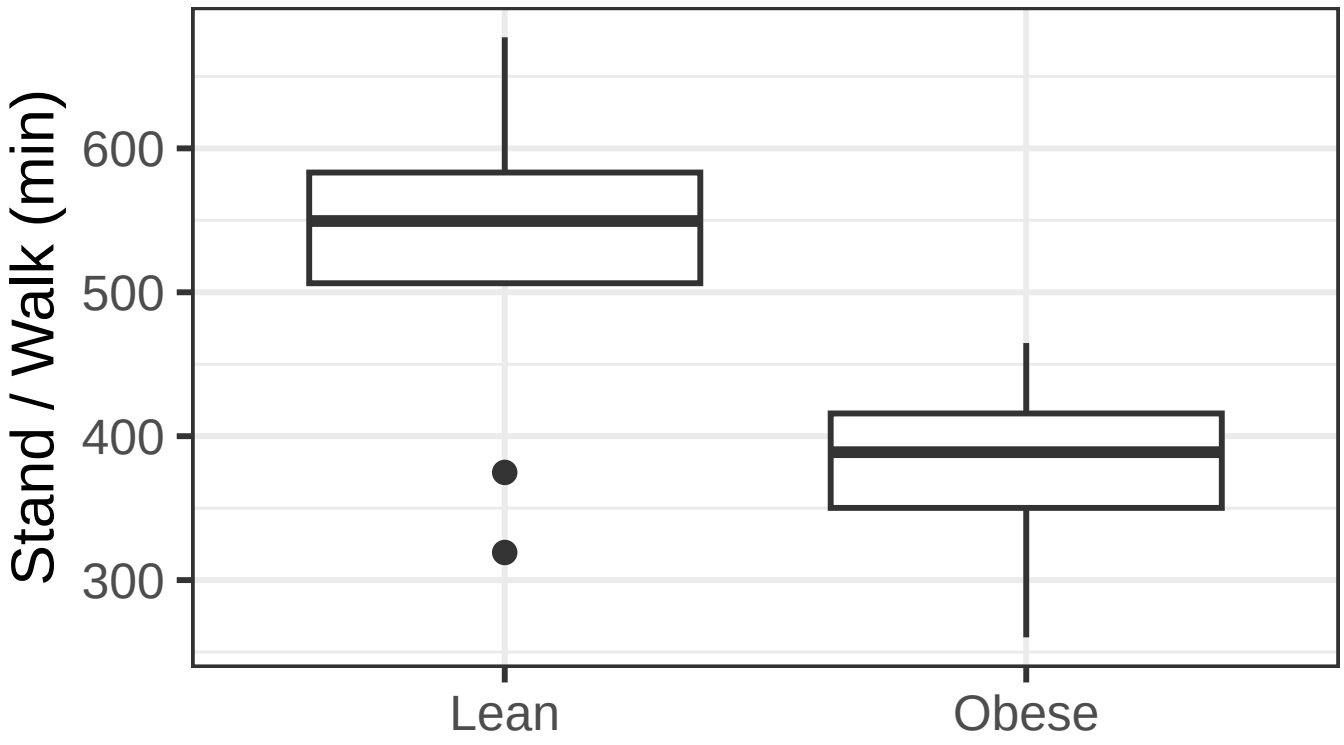
where t_C is the critical value with area C between $-t_C$ and t_C under the t-density with df degrees of freedom.

Example: Daily activity and Obesity.

- In one study, James Levine and collaborators at the Mayo Clinic investigated the link between obesity and daily activity[1].
- In this study, the researchers selected 20 individuals to track their daily activity. They selected 10 individuals who are lean, and 10 that are mildly obese (but still healthy). Sensors were attached to the subjects for 10 days, and recorded their daily movement patterns. The table below shows the number of minutes per data that the subjects spent standing or walking, sitting, and laying down.

[1] Levine, J.A., et al. 2005. Interindividual variation in posture allocation: possible role in human obesity. *Science, 307(5709)*, pp. 584-586

Plotting the Data



Looking at the data, we do not see any violations of the conditions for performing a two-sample t-procedure. The figure also shows a difference between the two groups; however, there are only 10 observations per group. How sure are we that the difference we see is not just due to random chance?

Two-sample t-procedures

To do this, we first calculate the sample mean and standard deviations:

Health	$\bar{X}$	s
Lean	525.751	107.121
Obese	373.269	67.498

Degrees of Freedom (df)

Using this data, we can calculate the degrees of freedom (df) using the formula provided above.

$$
\begin{aligned}
\text{df} &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} \\
&= \frac{\left(\frac{107.121^2}{10} + \frac{67.498^2}{10}\right)^2}{\frac{1}{9}\left(\frac{107.121^2}{10}\right)^2 + \frac{1}{9}\left(\frac{67.498^2}{10}\right)^2} \\
&\approx 15
\end{aligned}
$$

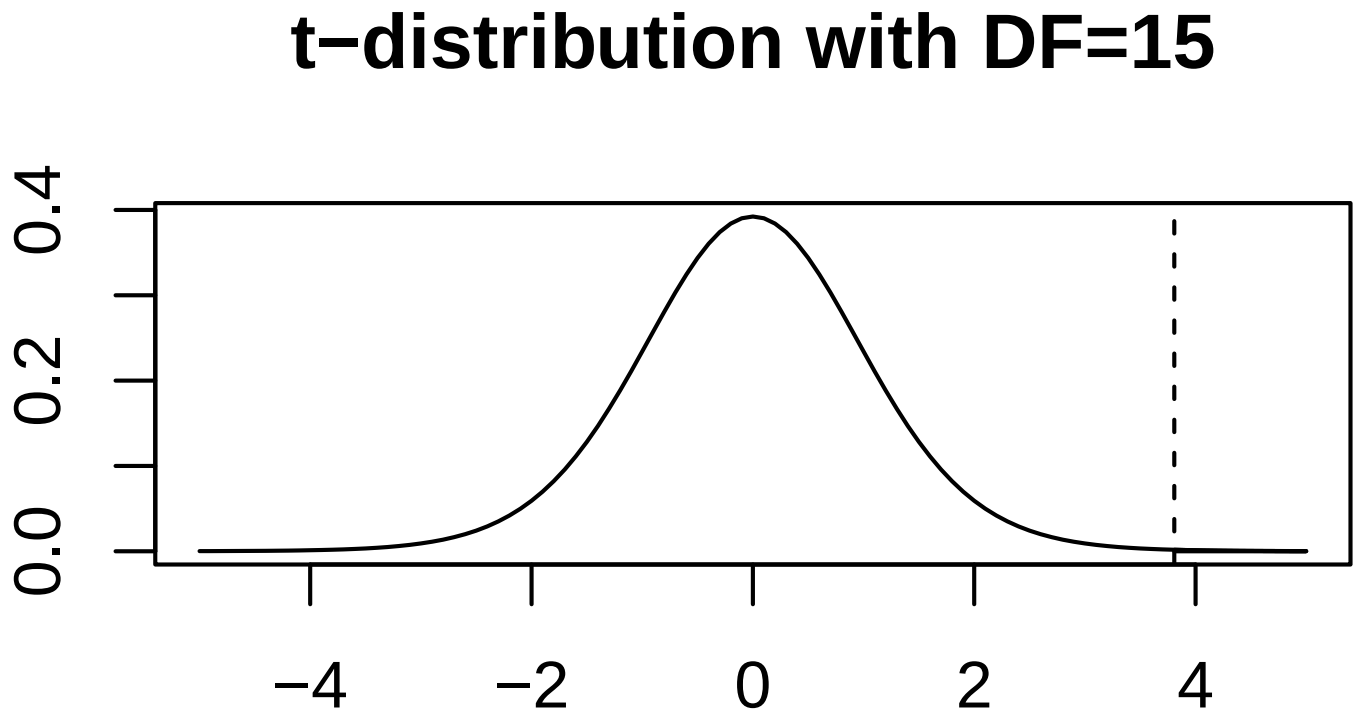
t-statistic

Now using the null-hypothesis that $\mu_1 = \mu_2$, we can calculate the t-statistic as:

$$
\begin{aligned}
t &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
&= \frac{525.751 - 373.269}{\sqrt{\frac{107.121^2}{10} + \frac{67.498^2}{10}}} \\
&= 3.808
\end{aligned}
$$

P-value

Below we plot the calculated t value compared to the theoretical sampling distribution under the null hypothesis:



As we can see, there is a very small chance that we would observe something as extreme as we did if the means were equal.

P-value (cont.)

We can calculate this probability, called the P-value, which is the area to the right of the dashed line (if a one-sided test), or twice this area (if a two-sided test).

The corresponding P-value can be calculated using software, in which case we get:

$$P\text{-value} = 0.0008$$

Using a t-table

Using a t-statistic table from the textbook, we can approximate the P-value by finding the lower and upper bounds of the value:

DF = 15		
t-statistic	3.733	4.073
One-Sided P	0.001	0.005

Our t-statistic (3.808) was calculated to be between 3.733 and 4.072, meaning the P-value must be between 0.001 and 0.0005. While the table doesn't provide us with the exact value, this is not critical for our conclusions. What relly matters is the order of magnitude of the P-value, not it's exact value.

Conclusions

What we have found is a statistically significant difference between the two samples.

Limitations:

- The findings come from an observational study, meaning we don't have a causal interpretation.
- We don't know for certain that $\mu_{\text{lean}} > \mu_{\text{obese}}$; we only have found evidence of this with our specific sample.

Confidence Interval

A confidence interval for $\mu_{\text{lean}} - \mu_{\text{obese}}$ is a better way to assess *how much* do lean individuals have higher activity rates than mildly obese individuals.

We want to obtain a 95% confidence interval for the difference in mean time spent walking or standing between the two populations. Recall from our previous calculations, we have the following estimates:

- $\bar{X}_{\text{Lean}} - \bar{X}_{\text{obese}} = 152.482$
- $\text{SE} = 40.039$
- $\text{DF} = 15.17 \approx 15.$

Calculating the critical value

For confidence intervals, we need to calculate the critical value for a t-distribution with the correct degrees of freedom. If we approximate $\text{df} = 15$, the critical value for a 95% confidence interval is

$$t_{95\%} = 2.131,$$

and the confidence interval can be evaluated as:

$$\text{CI:} \quad (\bar{X}_1 - \bar{X}_2) \pm t_{95\%} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{1}$$

$$\text{CI:} \quad 152.482 \pm 2.131 \times 40.039 \tag{2}$$

$$\text{CI:} \quad (67.159, 237.805) \tag{3}$$

Interpretation

The difference in mean daily standing/walking times between the two groups is estimated to be on the order of one to four hours. The confidence interval is very wide because (1) the populations are small, and (2) the sample standard deviations are large. The variability of individuals is something we cannot control, but we could obtain tighter confidence intervals if we sampled more data.

Example: Transgenic Chickens

Infection of chickens with the avian flu is a threat to poultry production and human health. A research team created transgenic chickens that are resistant to infection, but would like to know if the modification also affects chicken sizes. The researchers compared the hatching weights (in grams) of 45 transgenic chickens to 54 standard chickens of the same breed.

Data

Transgenic

38.8, 39, 39.7, 40, 40.8, 40.9, 41, 41, 41, 42.5, 42.6, 43, 43, 43.4, 43.5, 43.5, 43.8, 44.4, 44.7, 44.7, 44.7, 45.3, 45.7, 45.8, 46.4, 46.5, 46.6, 46.7, 46.7, 46.8, 46.9, 47.1, 47.1, 47.1, 47.3, 47.6, 47.7, 48.1, 48.3, 49.3, 49.3, 49.8, 50.3, 50.9, 52.1

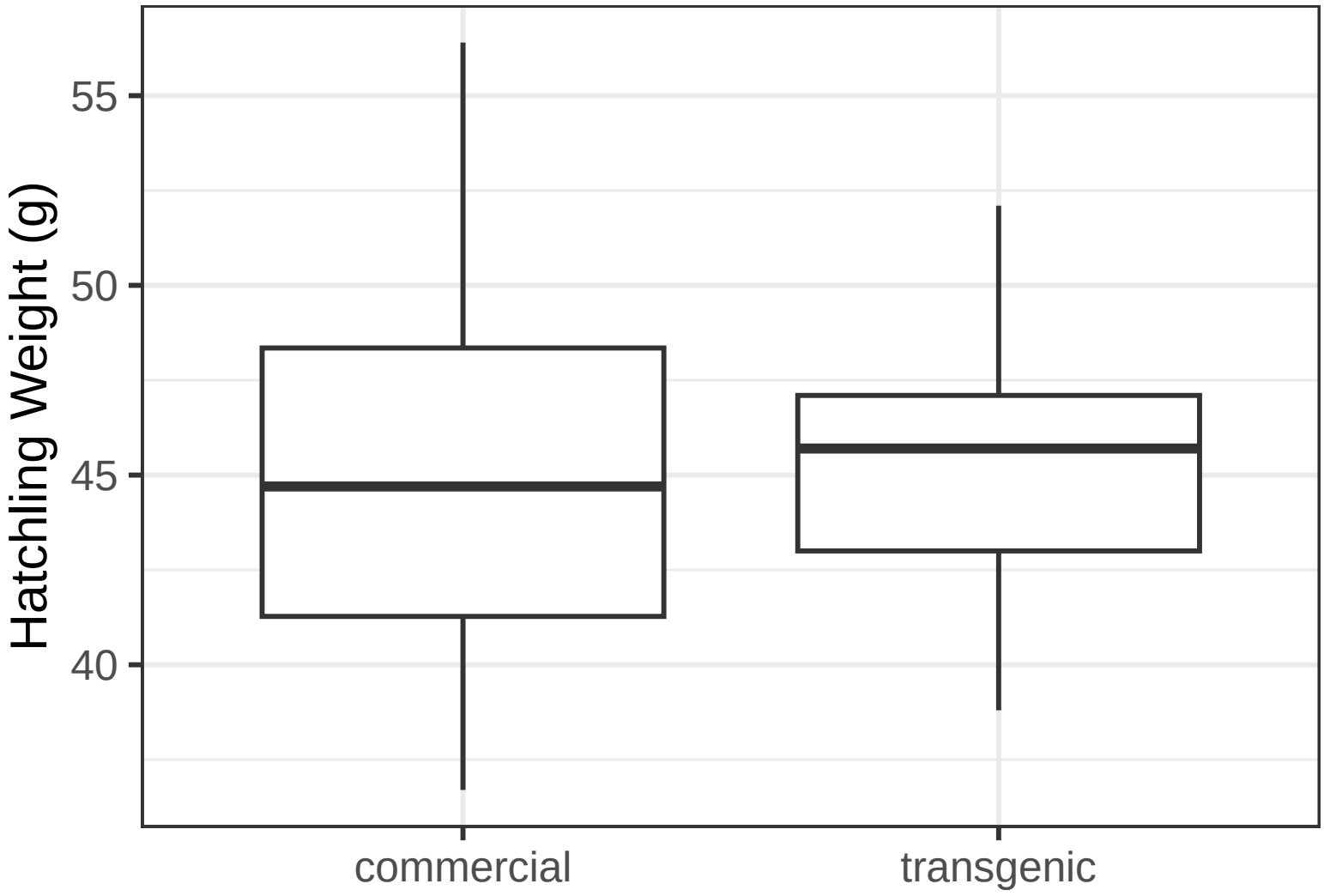
Commercial

36.7, 37.1, 38.9, 39.5, 39.5, 39.8, 40, 40.2, 40.3, 40.5, 40.5, 40.7, 41.1, 41.2, 41.5, 41.5, 41.6, 41.6, 41.7, 42.4, 43.1, 43.3, 43.3, 43.4, 43.7, 44.1, 44.2, 45.2, 45.3, 45.4, 46, 46.1, 46.4, 46.6, 46.6, 46.9, 47.3, 47.5, 48.1, 48.2, 48.4, 48.6, 49, 49.1, 49.3, 49.6, 50.1, 50.2, 50.4, 50.6, 52.2, 53, 55.5, 56.4

Goal

As before, we are interested in testing the hypothesis $H_0 : \mu_{\text{transgenic}} \neq \mu_{\text{commercial}}$, and finding a 95% confidence interval for $\mu_{\text{transgenic}} - \mu_{\text{commercial}}$

Plot



Sample Statistics

The first step is to calculate the sample statistics:

Type	$\bar{X}$	s
Commercial	44.989	4.569
Transgenic	45.142	3.321

t-statistic

The t-statistic is calculated as

$$
\begin{aligned}
t &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\
&= \frac{45.14 - 44.99}{\sqrt{\frac{3.32^2}{45} + \frac{4.57^2}{54}}} = 0.19
\end{aligned}
$$

Degrees of Freedom (df)

The degrees of freedom are calculated as

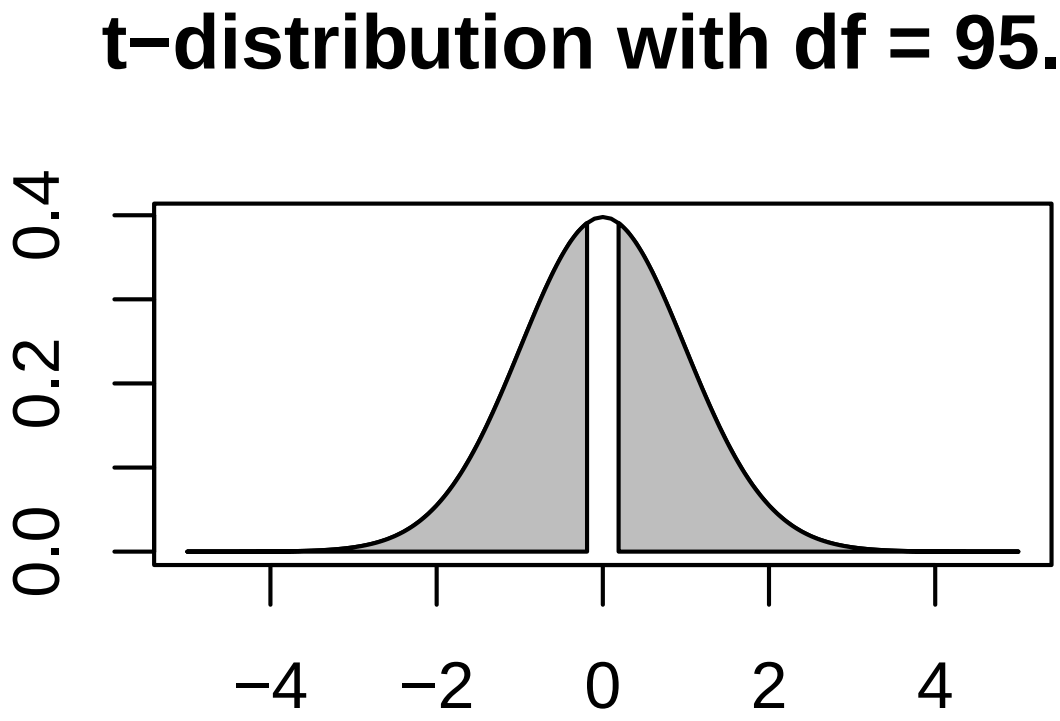
$$\begin{aligned}
\text{df} &= \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \\
&= \frac{\left(\frac{3.32^2}{45} + \frac{4.57^2}{54}\right)^2}{\frac{1}{44}\left(\frac{3.32^2}{45}\right)^2 + \frac{1}{53}\left(\frac{4.57^2}{54}\right)^2} = 95.3
\end{aligned}$$

P-value

In this case, the t-statistic we calculated is very small. Using statistical software, the two-sided P-value from this sampling distribution is calcated to be:

$$P\text{-Value} = 0.85,$$

P-value (cont.)



This result shows that there is not strong evidence that there is a difference between mean weights from the two distinct populations.

Confidence Interval

The corresponding 95% confidence interval can be calculated using a critical value of

$$t_{95\%} = 1.985,$$

which was calculated using statistical software. The 95% confidence interval is therefore given by

$$\text{CI:} \quad (\bar{X}_1 - \bar{X}_2) \pm t_{95\%} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{4}$$

$$\text{CI:} \quad 0.15 \pm 1.985 \times 0.795 \tag{5}$$

$$\text{CI:} \quad (-1.42, 1.73) \tag{6}$$

Therefore $\mu_{\text{transgenic}} - \mu_{\text{commercial}}$ could be negative, positive, or zero.

Pooled two-sample t-proceedures

In the *t*-procedures discussed above, we allow for the estimation of distinct population standard deviations, σ_1 and σ_2. This is known as *unpooled* *t*-procedures, or sometimes the *Welch's* *t*-procedure.

Pooled standard deviation estimates

An alternative approach would be to assume that the populations have the same standard deviation, $\sigma_1 = \sigma_2 = \sigma$. This is known as *pooled* t-procedures, and in this case we combine the sample standard deviations for each group (s_1, s_2) in order to get a single estimate of the population standard deviation s_p:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

Pooled t-statistic

Using the equal (or pooled) estimate of the standard deviation, the t-statistic can be calculated as:

$$t_{\text{pooled}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

with degrees of freedom:

$$\text{df} = n_1 + n_2 - 2.$$

Which Test should I use?

Most software packages offer a choice of two-sample t-statistics using either *unequal* or *equal* population variances. Which test should we chose to use? Here are a few considerations:

- The *unequal* test is valid whether or not the population variances are equal, whereas the *equal* test can be unreliable if the variances are not actually equal.
- The *equal* variance procedure is more simple to calculate by hand, but we typically have software available to calculate the *unequal* version.
- Simulation studies suggest that if the population variances are not exactly equal (as they often are not in real datasets), then the *unequal* test gives more reliable estimates.
- If the sample size is small, *and* the variances of the two groups are equal, the *equal* variance procedures can be better. **However**, in practice we cannot know if the population variances are equal.