t-tests

Nicholas Kortessis

t-tests: Comparisons of means

t-tests are a classic tool in the biostatistician's toolkit. They are called t-tests because they rely on the t-distribution. Remember, t-distributions represent the sampling distribution of the mean of a sample of normally distributed outcomes.

Stated simply, if your data are normally distributed (or somewhat close to it), the sample mean follows a t-distribution.

This idea is leveraged to compare the sample mean with some other mean and evaluated whether any difference is consistent with sampling variability.

Remember that when we were calculating the confidence interval for the sample mean $\bar{X}$, we used the t-statistic

$$t_{\text{stat}} = \frac{\bar{X} - \mu}{SE_{\bar{X}}}.$$

This statistic measures how much the sample mean differs from the population mean in units of standard deviation.

To use this for hypothesis testing, we need to do the same thing but instead of including *the population mean,* μ, we will put in an assumed value that represents the hypothesis.

Let's start with a very simple example, the soil pH example from lab with the "worms" dataset. The setup is that there are different fields where the researcher wants to know something about the conditions that influence how abundant earthworms are. Among the things measured by the researchers is the pH of the soil of each field. All the fields in the study are grasslands. Grasslands have a typical pH. We want to ask the question of whether these fields represent 'typical' grasslands according to their pH.

This is a great question for a t-test. Let X_i be the pH of the ith grassland in the study. The average pH of a grassland is 6.5. We can evaluate whether the soil pH of all these grasslands is consistent with this expectation. Let's set the hypothesis that the mean soil pH is 6.5. We do this by writing

$$H_0 : \mu = 6.5$$

$$H_A : \mu \neq 6.5.$$

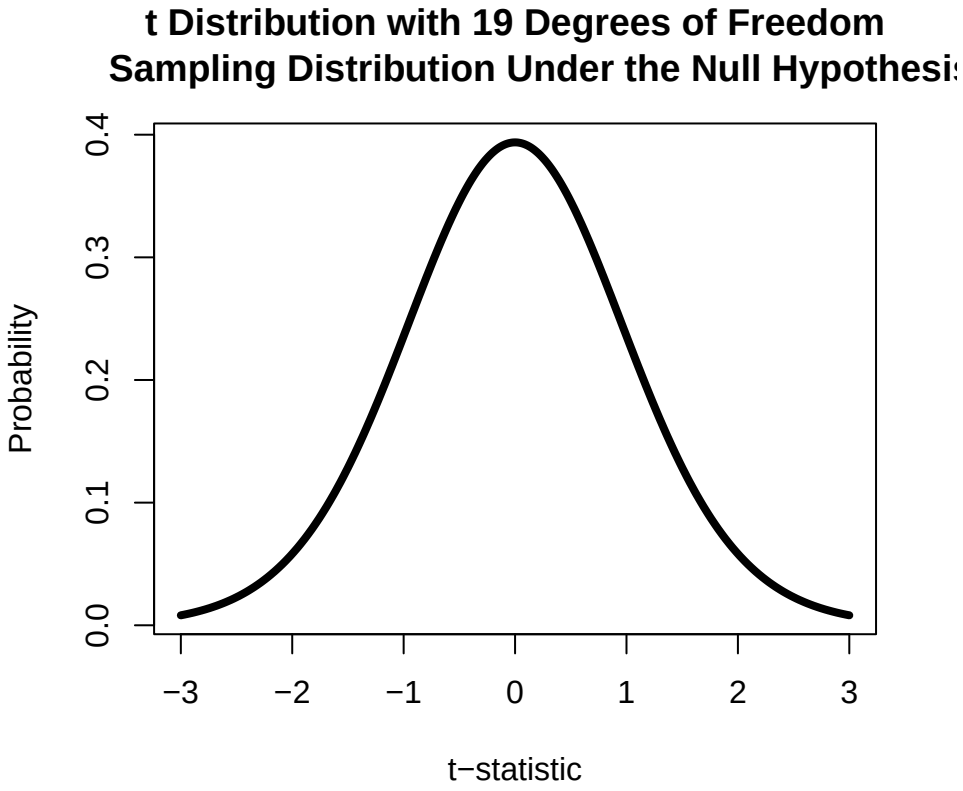
Thus, the null hypothesis is that the average soil pH is 6.5 and the alternative is that it is something else.

The approach of a t-test is to say that, if this null hypothesis is true, then the sampling distribution of the sample mean pH, $\bar{X}$, can be characterized by the t-statistic

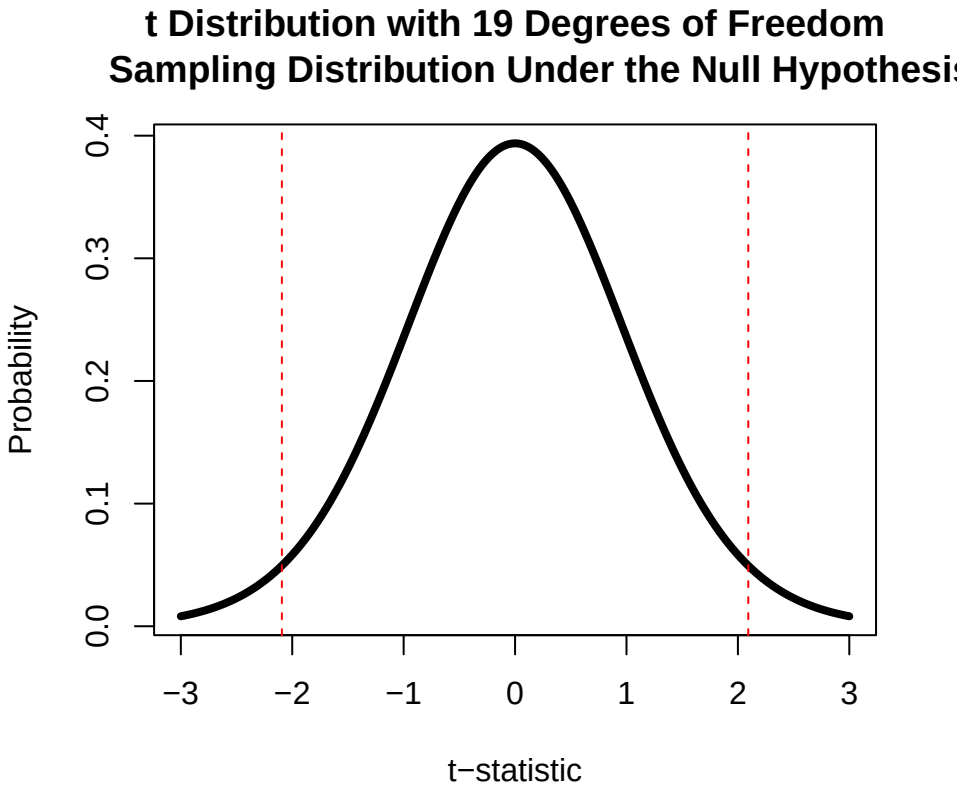
$$t_{\text{stat}} = \frac{\bar{X} - 6.5}{SE_{\bar{X}}}.$$

The worms dataset has $n = 20$ fields (i.e., statistical individuals), and so this sampling distribution has $n - 1 = 19$ degrees of freedom. Here is the data set.

Here is the sampling distribution of the t-statistic under the null hypothesis.



With an α value of 0.05, we can find the cutoff for the rejection region using the quantile function for the t-distribution. (In R, the function for the lower cutoff is `qt(0.025, df = 19)` and for the upper cutoff is `qt(0.975, df = 19)`). Here it is plotted.

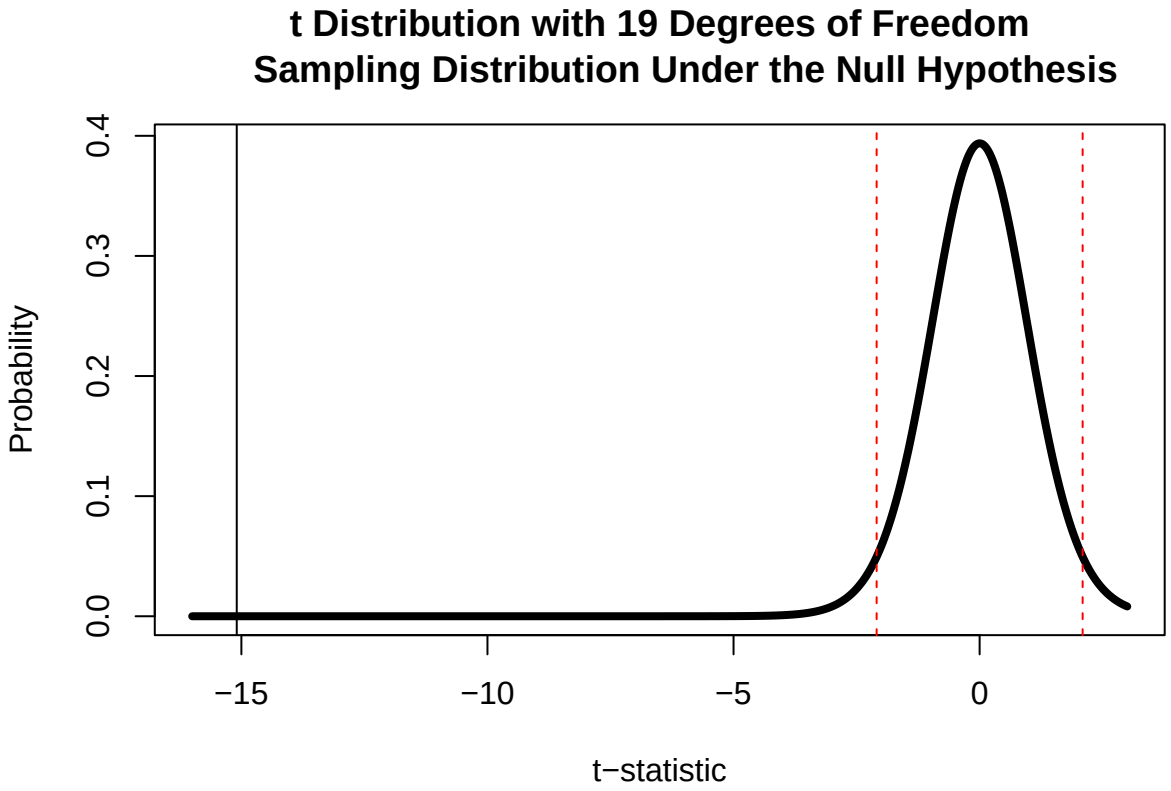


The cutoff values in this case are -2.093 and 2.093. If the observed t-statistic from our soil pH data is within the range [-2.093,2.093], we fail to reject the null hypothesis that the average soil pH is 6.5. If it is outside this region, then we reject the null hypothesis.

The sample mean soil pH is $\bar{X} = 4.55$ and the sample standard deviation is $s = 0.576$. This means that the standard error of the mean is $SE_{\bar{X}} = s/\sqrt{n} = 0.576/\sqrt{20} = 0.129$. As such, the t-statistic for our sample of fields is

$$t = \frac{\bar{X} - \mu_0}{SE_{\bar{X}}} = \frac{4.55 - 6.5}{0.129} = -15.1$$

That's **way** below the cutoff, and very clearly in the rejection region. This t-value indicates that the mean soil pH is 15.1 standards errors lower than the hypothesis. That's extremely unlikely if the null hypothesis is true. As such, we reject the hypothesis that the average soil pH for the fields under study is 6.5. Here is what it looks like on the sampling distribution.



Now we can evaluate the probability of a more extreme outcome. That is, a more extreme test statistic. That probability is

$$p\text{-value} = \Pr(t < -15.1 \text{ or } t > 15.1).$$

We want to know the probability that t is smaller than our calculated statistic, but also that it is larger than an equivalently large t-statistic on the other side of the hypothesis, i.e., $t > 15.1$. This is because our hypothesis only states that the mean of the population is 6.5. pH very far below 6.5 counts as extreme but also pH very far **above** 6.5 also counts as extreme.

We can calculate this p-value using the cumulative density functions for the t-distribution as follows.

$$p\text{-value} = \Pr(t < -15.1 \text{ or } t > 15.1)$$
$$p\text{-value} = \Pr(t < -15.1) + \Pr(t > 15.1)$$
$$p\text{-value} = \Pr(t < -15.1) + 1 - \Pr(t < 15.1)$$

These two probabilities are cumulative probabilities of the t-distribution. The first we can find with the code `pt(-15.1, df = 20-1)` and the second we can find with the code `pt(15.1, df = 20-1)`. These two together give the p-value of

```
pt(-15.1, df = 19) + 1 - pt(15.1, df = 19)
```

```
## [1] 4.896528e-12
```

This tells us that we are very unlikely to have seen this result in the null hypothesis is true. All these pieces of information point that we should be strongly skeptical of the null hypothesis for this data.

Kinds of t-tests, their associated statistics, and their sampling distributions

t-tests come in three forms

1. One-sample t-test

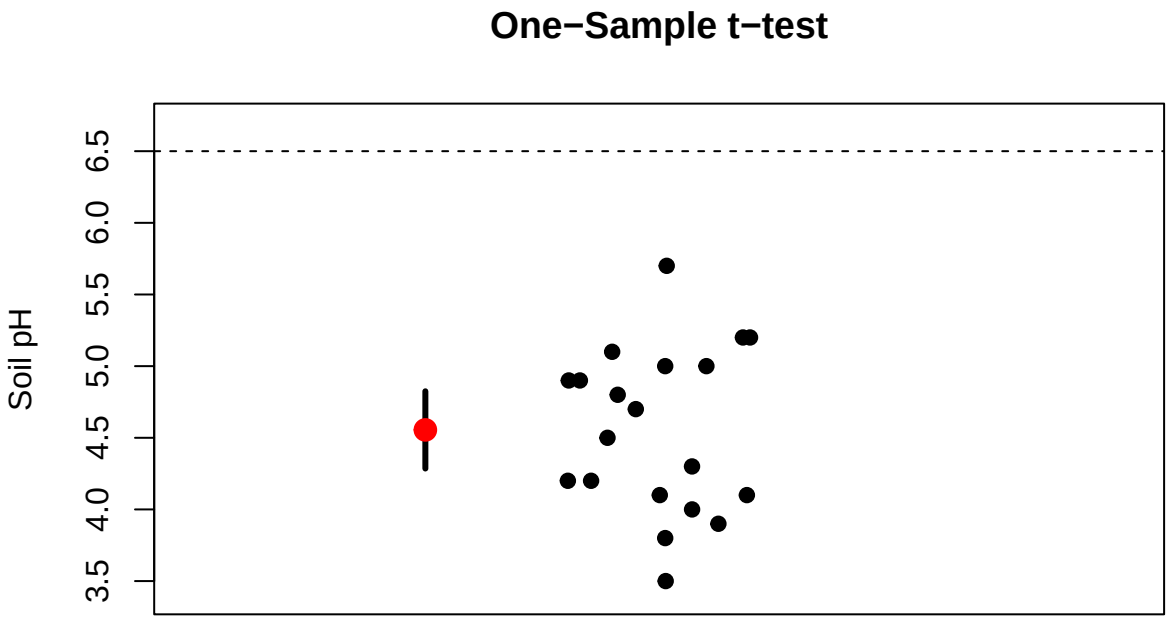
   One-sample t-tests are used when you want to know if the mean of a specific sample is consistent with a hypothesized value. The example of soil pH is an example of a one-sample t-test.

   The t-statistic is just

$$t = \frac{\bar{X} - \mu_0}{SE_{\bar{X}}}$$

   and this t-statistic follows a t-distribution with degrees of freedom equal to one less than the number of individuals in the group (i.e., $df = n - 1$).

   A good way to visualize a one sample t-test is to use a stripchart in relation to the null hypothesis and showing the mean of the sample and the confidence interval of the mean. Here is an example with the soil pH data in relation to the hypothesis that mean soil pH is 6.5.



   This figure shows the soil pH for each field (black dots), the mean soil pH for the sample (the red dot) and the associated 95% confidence interval. The hypothesis of 6.5 soil pH is given by the dashed line. You can see plain as day that there is little chance the actual mean of the fields is 6.5.

2. Two-sample t-test

   Two sample t-tests are when you compare the means of two different groups. While it is possible to make this comparison with respect to some hypothesized difference, the cases where that is a plausible are limited and rarely done in practice. Most commonly, Two-sample t-tests have a null hypothesis where the difference in the means of the two groups is zero. The two-sample t-test works just like a one-sample t-test because the distribution of the difference in means is also normally distributed. As such, the difference of the means scaled by their standard error is also distributed as a t-distribution.

   The specific t-statistic for a two-sample t-test is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_{01} - \mu_{02})}{SE_{\bar{X}_1 - \bar{X}_2}}$$

where $\bar{X}_1$ is the sample mean of the first group, $\bar{X}_2$ is the sample mean of the second group, $\mu_{01} - \mu_{02}$ is the hypothesized difference in the population means of the two groups, and $SE_{\bar{X}_1 - \bar{X}_2}$ is the standard error of the difference in the sample means of the two groups. This is identical to the one-sample t-test except now we are interested in modeling *differences between groups*. Our measure of differences is $\bar{X}_1 - \bar{X}_2$. our hypothesized difference is $\mu_{01} - \mu_{02}$, and just like how we can represent how variable a sample mean is from sample to sample, we can also characterize how variable the *difference in the sample means are* from sample to sample. Again, the null hypothesis is almost always taken as $\mu_{01} - \mu_{02} = 0$, and so that term if often omitted in the equation for a two-sample t-test. But it's all the same.

Luckily the standard error of the difference in the means is pretty easy to calculate. It is

$$SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{SE_{\bar{X}_1}^2 + SE_{\bar{X}_2}^2}$$

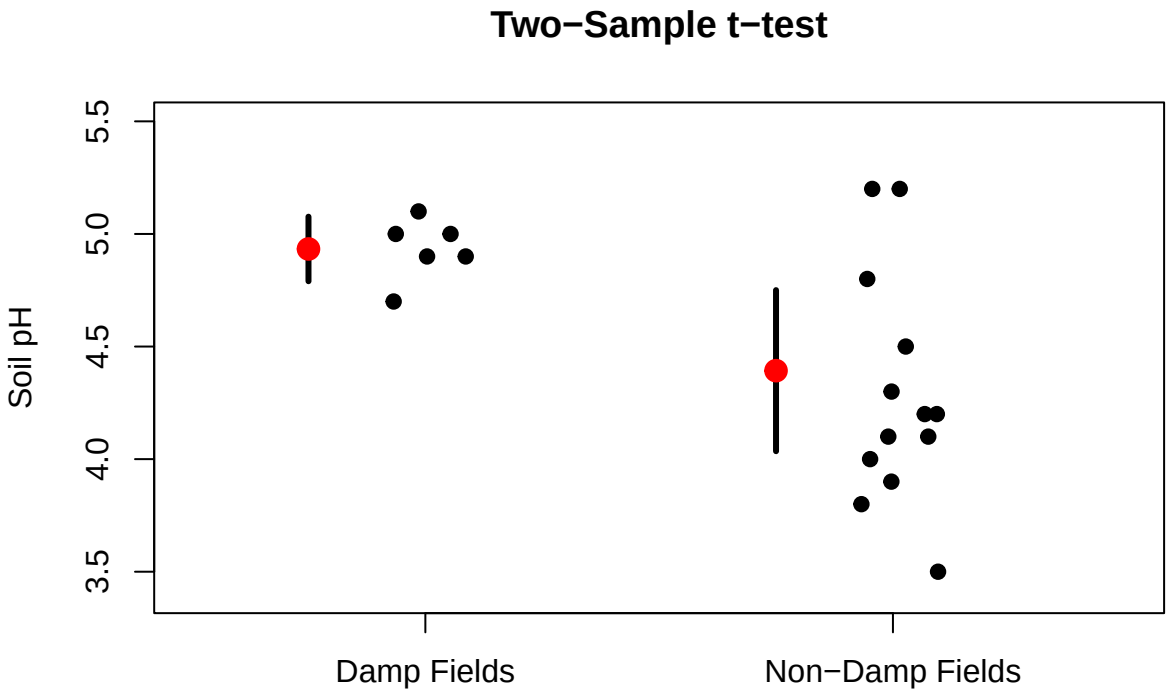
which is just the square root of the variances of the sampling distributions of each sample mean. This gets at why a two-sample t-test is not the sample as a one-sample t-test when comparing against the mean of another sample. There is uncertainty in the estimate of the mean of both samples (as quantified by $SE_{\bar{X}_1}$ and $SE_{\bar{X}_2}$) that combine together to lead to greater uncertainty than we would expect with just one group.

You will find many equations for the standard error of the difference between means on the internet and in the book. All are equivalent to this equation, as this is the most general. Those other equations work in some simple cases, such as equal variances between the groups and equal sample sizes. No need to worry about those here.

The t-statistic for a two-sample t-test follows a t-distribution with degrees of freedom equal to $n_1 + n_2 - 2$, which is the sum of the degrees of freedom for each group.

Some examples of when two-sample t-tests are used is when you have two experimental treatments (typically control and treatment) and you want to see if there is an average effect of treatment. Another example comes from the worms study. Does soil pH differ between damp and non-damp fields? That can be answered by a two-sample t-test because soil pH is normally distributed and because this is a question about the relationship between the averages of two populations: damp fields and non-damp fields.

A way to visualize the two-sample t-test is by plotting data points, means, and associated confidence for each of the two groups on the same graph. The example below shows this for soil pH in damp and non-damp fields.



3. Paired t-test

Paired t-tests work like a mix of one-sample and two-sample t-tests. They are used when the same statistical individuals are used more than once to create what looks like two samples. Let's start with an example. Imagine you are doing a study of soil pH in these fields, and you want to know the effect of worms on soil pH. One way to study the effect of worms is to measure soil pH with the earthworms there and then to remove the earthworms and then re-measure soil pH sometime later.

By running this experiment, soil pH is measured **twice from the same field**: once with the earthworms and once without. As such, the two measurements from the same field are paired. Statistically, the important thing is that they are not independent, which is to say that these two samples have NOT been taken randomly. But a way to deal with this non-independence is to run a one-sample t-test on **the differences of the paired values.**

For each statistical individual, we calculate

$$\Delta X_i = X_{i,\text{pre}} - X_{i,\text{post}}$$

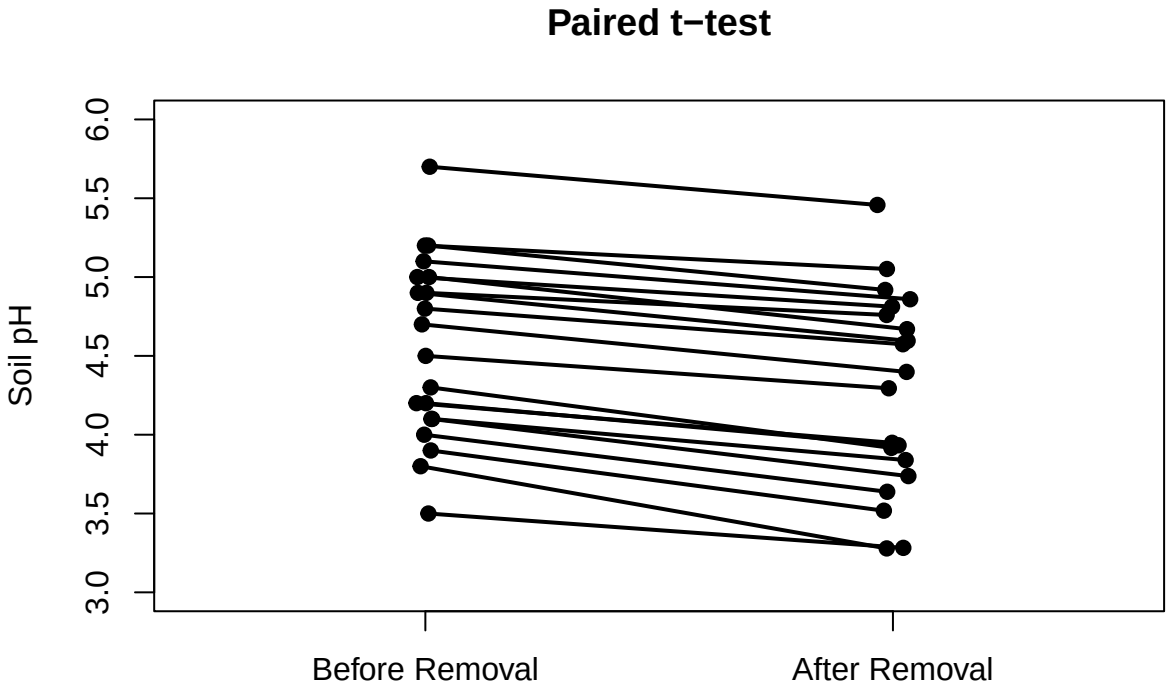
where $X_{i,\text{pre}}$ is soil pH before removing worms in field i, and $X_{i,\text{post}}$ is soil pH after removing earthworms in field i. Thus, ΔX_i is **the effect of removing worms** in field i.

The t-statistic for this kind of paired data is

$$t = \frac{\overline{\Delta X} - \mu_{0\Delta}}{SE_{\overline{\Delta X}}}.$$

In this equation, $\overline{\Delta X}$ is the average of the paired differences in the sample, $\mu_{0,\Delta}$ is the hypothesized average difference of pairs in the population, and $SE_{\overline{\Delta X}}$ is the standard error of the average differences. As is the case with the two-sample t-test, we typically assume $\mu_{0,\Delta} = 0$, representing a straw man assumption that there is no effect of removing earthworms (or any other factor the links pairs of measurements together). This is exactly identical to running a one-sample t-test but where the sample values are just the differences in the pairs.

To visualize paired t-tests, plot the data as if it were a two-sample t-test but join together paired points with a line. We don't have data to visualize this from the worms dataset from class, but we can make up something. Here is an example.



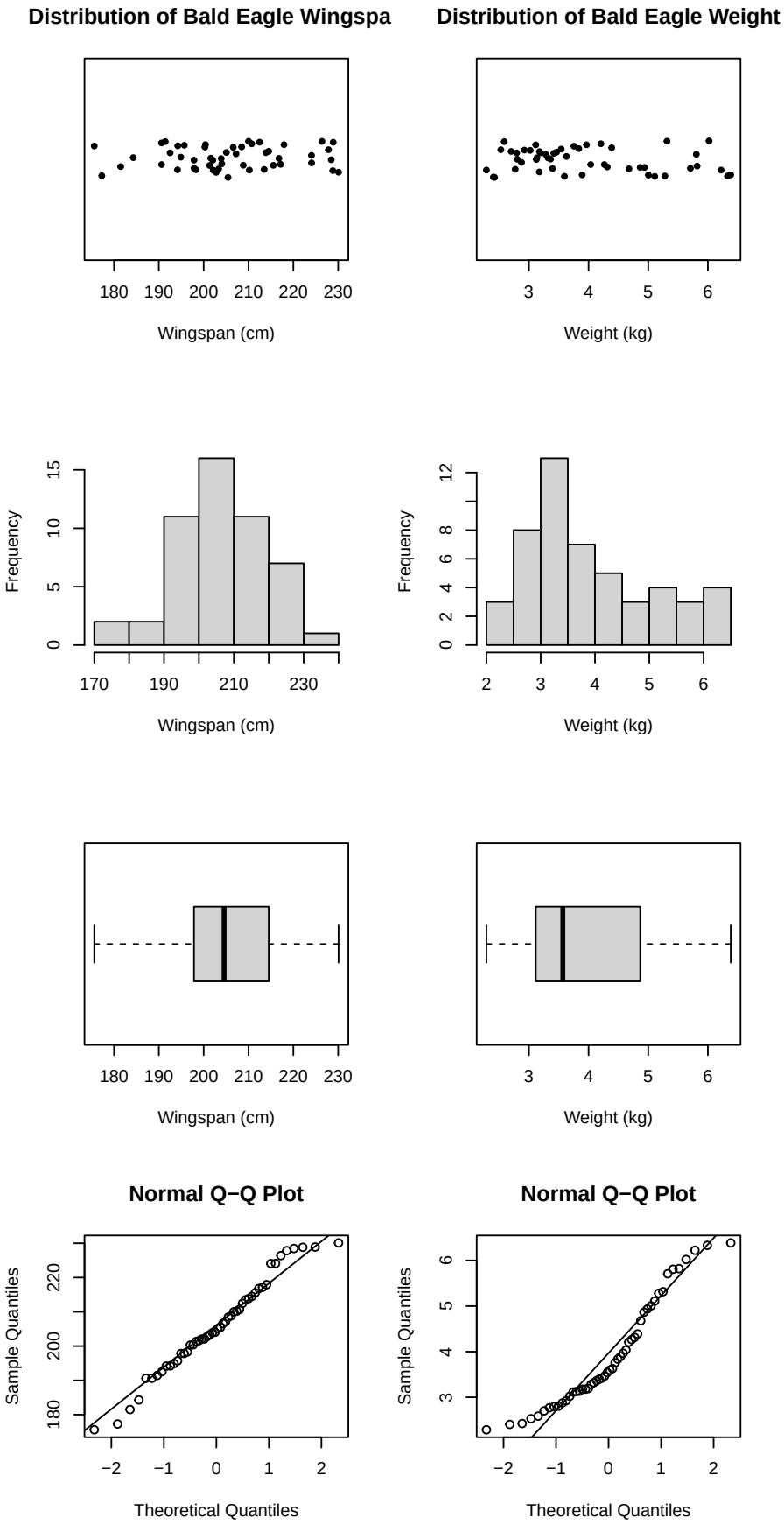
You can see now from this figure which points in correspond to the same field because they are paired by a line. Here, the question is about whether the slopes of the lines are, on average, different from zero. The slopes represent the difference between paired points. The slopes have an average ($\overline{\Delta X}$) and we can estimate the variability in the average slope ($SE_{\overline{\Delta X}}$). Here, there appears to be a moderate, but consistent negative effect of removing earthworms on soil pH.

What do t-tests assume?

t-tests follow all the assumptions that go along with a t-distribution. The only real assumptions is that the individual groups are normally distributed.

This means that when you look within your data, the groups should be normally distributed. It's that simple. To evaluate consistency of the data with a t-test model, then you need to look at the distribution of the data within each group. You can do this with a histogram, a stripchart, a boxplot, or a density plot. You can also resort to using a Q-Q plot to help!

Here is an example with wingspans and weights of bald eagles.


Distribution of Bald Eagle Wingspa
Distribution of Bald Eagle Weight
180 190 200 210 220 230
Wingspan (cm)
3 4 5 6
Weight (kg)
Frequency
0 5 10 15
170 190 210 230
Wingspan (cm)
Frequency
0 2 4 6 8 12
2 3 4 5 6
Weight (kg)
180 190 200 210 220 230
Wingspan (cm)
3 4 5 6
Weight (kg)
Normal Q−Q Plot
Normal Q−Q Plot
Sample Quantiles
180 200 220
−2 −1 0 1 2
Theoretical Quantiles
Sample Quantiles
3 4 5 6
−2 −1 0 1 2
Theoretical Quantiles

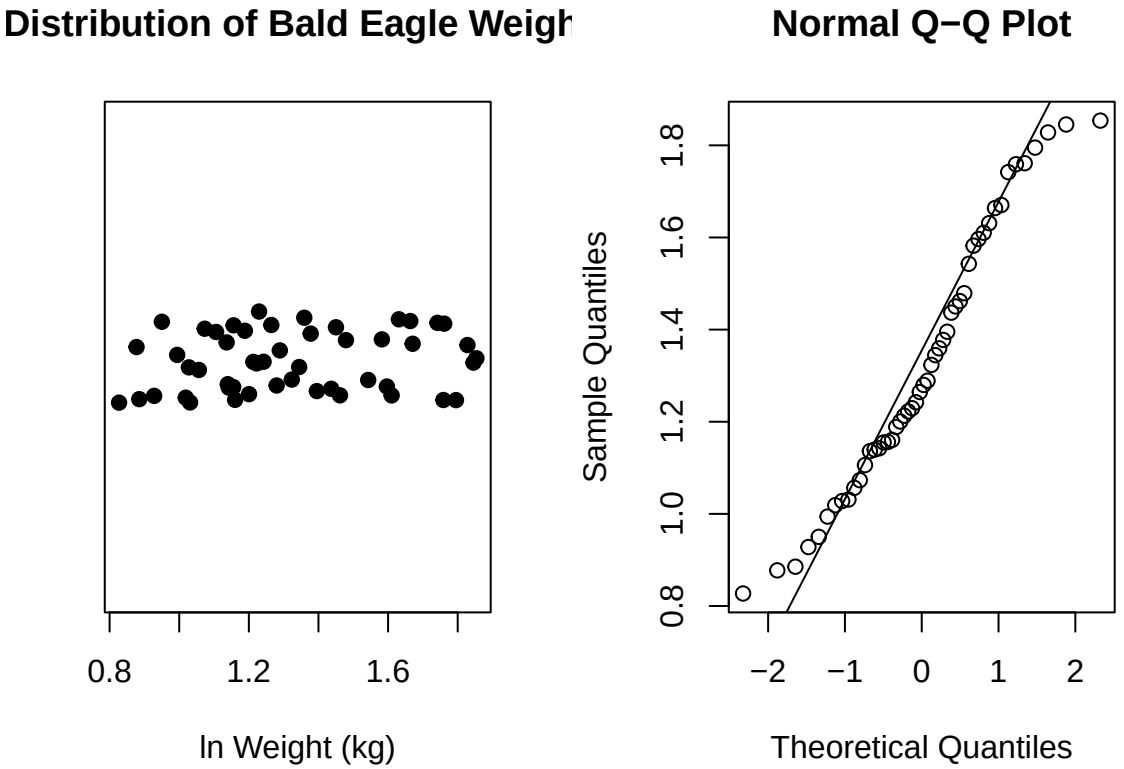
There is some evidence that wingspan can be modeled with a normal distribution. It's unimodal, is pretty symmetrical, and doesn't seem to have many extreme outcomes. The Q-Q plot also shows pretty good correspondence between the points and the line.

We should look at the data on eagle weight a bit more skeptically. There is some sign that it is isn't symmetric and there is obvious curvature to the Q-Q plot.

Can you ever do t-tests without normally distributed populations?

The fundamental thing to appreciate is that **the data points that go into the analysis can be modeled as normally distributed.** One way to do this is to try transforming the data. One transformation that is common is to take the logarithm of the data (any logarithm will do; different fields have different conventions and some are easier to interpret than others). Others include a square root transformation and a box-cox transformation.

Applying the natural logarithm to the bald eagle weight does a good job of making the data more normally distributed.



One the data are transformed and meet the assumptions of normal model, then t-tests can be safely applied. However, remember to think about the interpretation of the data and test results apply to the transformed data. Often, you want to present the results on the original scale, which is often easier to interpret. Sometimes it can be more difficult to present. In any case, you'll need to balance ease of interpretation with ease of visualization. Some creativity and thought is required.

t-test effect sizes

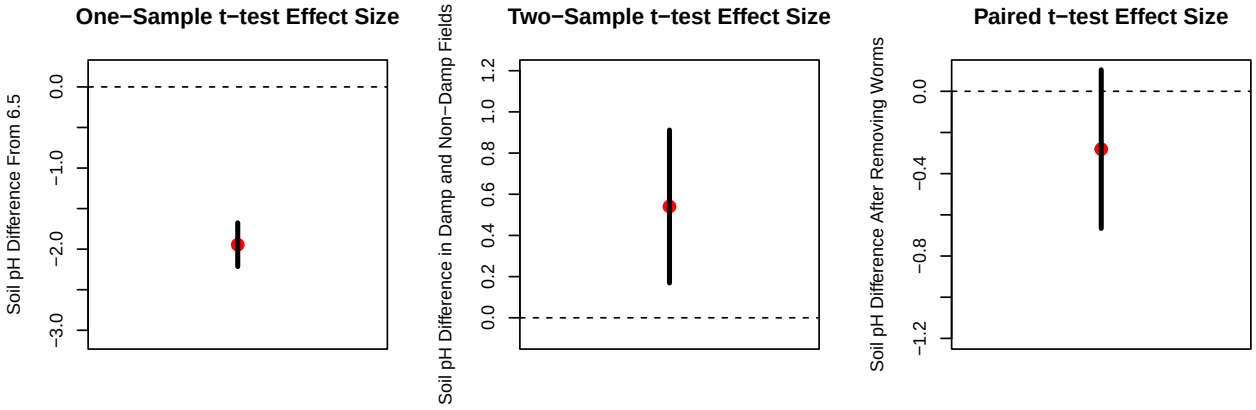
All t-tests are united by thinking about them in terms of effect sizes.

The effect size in each case is the estimated difference between the sample mean and the hypothesis. For one-sample t-tests, the effect size is $\bar{X} - \mu_0$. For a two-sample t-test where the null is no difference between the groups, the effect size is $\bar{X}_1 - \bar{X}_2$. For a paired t-test, the effect size is $\overline{\Delta X}$.

In each case, the effect sizes can be estimated using confidence intervals. These confidence intervals are all determined by the t-distribution and shaped by the uncertainty in the effect as quantified by the appropriate standard error. The table below shows the measure of uncertainty and the effect size for each type of t-test.

	One-Sample t-test	Two-Sample t-test	Paired t-test
Effect Size	$\bar{X} - \mu_0$	$\bar{X}_1 - \bar{X}_2$	$\overline{\Delta X}$
Standard Error	$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$	$SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}$	$SE_{\overline{\Delta X}} = \frac{s_{\overline{\Delta X}}}{n}$

You can always determine the outcome of t-test by looking at a confidence interval of the effect size in each case. The figure below shows the graphs for the effect size and the 95% confidence interval of effect size in the three examples given above.



When we show the effect size and the 95% confidence interval, a t-test just evaluates whether the 95% confidence interval crosses the zero line. That's a very simple interpretation that unifies all three t-tests.

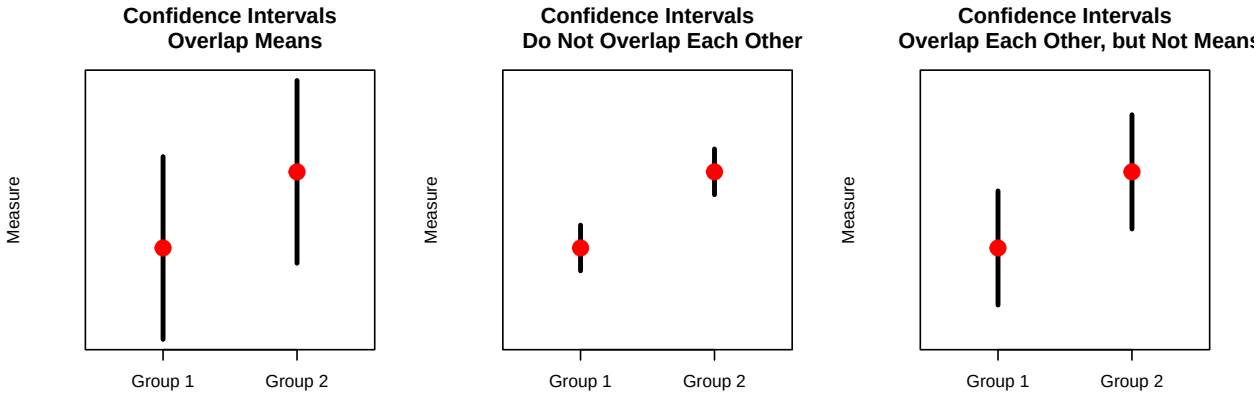
Determining statistical significance using confidence intervals

When you don't have effect sizes, can you use confidence intervals to evaluate statistical significance?

The answer is always "yes" for one-sample tests. Just ask when the confidence interval for the mean overlaps the hypothesized mean value.

The answer is always "no" for paired tests, unless you do like above and present the effect size. A plot of the means of groups with confidence intervals is not informative when there are paired individuals in the groups.

The answer is "sometimes yes" and "sometimes no" for two-sample tests. There are three possibilities, outlined in the three graphs below. The first (left panel) is where the confidence intervals overlap with the means. In that case, you can safely say the two groups are NOT statistically different. The second (middle panel) is where the two confidence intervals do not overlap each other. You can safely say in that case that the two groups ARE statistically different. The third (right panel) is where the confidence intervals overlap each other but not the means. In this case, you can't tell whether the groups are statistically different from one another. The test outcome can only be determined from the confidence interval of the effect size (as outlined above).



Interpreting the t-value

The t-statistic has three fundamental quantities: the effect size, the variability of the sample(s), and the sample size. You can see each of these in the t-statistic. For simplicity, let's look at the t-statistic in the example of a one-sample t-test (but these arguments apply to all t-tests). As a reminder, here it is

$$t = \frac{\bar{X} - \mu_0}{SE_{\bar{X}}}.$$

It includes the difference of the sample mean from the hypothesized mean divided by the standard error of the mean. Remember that the standard error includes two values, the sample standard deviation and the sample size because $SE_{\bar{X}} = s/\sqrt{n}$. We can rewrite the t-statistic to show each of these things as follows

$$t = \frac{(\bar{X} - \mu_0)}{s}\sqrt{n}.$$

This shows a **very important relationship.** t-values can be large (and so p-values will be small) for three different reasons:

1. Effect sizes $(\bar{X} - \mu_0)$ are large. This indicates a big effect compared to the hypothesis.
2. Population variability (s) is small. When the population has low variability, s in the denominator is small and so the t-statistic can be large. This is a sign of certainty in the effect size.
3. Sample sizes (n) are large. Even if effect sizes are moderate and population variability is large, large sample sizes can provide lots of evidence and so lead to large t-values (and small p-values).

Thus, there are multiple reasons that lead to statistical significance. Some have to do with the variability in the problem. Some have to do with the size of the experiment. Only one has to do with how big the effect is.

The size of the effect is usually related to **biological significance.** To illustrate this, imagine we do a paired t-test with a drug meant to improve human health. With a t-value or p-value alone, we can make a statement about the consistency of a drug trial with the hypothesis of no effect of the drug. However, with this information alone, we cannot say much about **how much the drug improves health.** That is because t-values include effect sizes but also include other factors. Even very small effects can lead to large t-values (and small p-values) if the sample size is large enough.

The key take-home point here is that **statistical significance is not the same as biological significance.** To clearly get at biological significance, show effect sizes. To get at statistical significance, show effect size confidence intervals! These two together give most of the information necessary to make an evaluation of the evidence for an effect and the relevance of an effect for biology.