3.2: T-Tests

Comparing Two Means

When we are comparing data from two groups, we often want to compare the mean values of the different groups to see if there are differences. We can do this in a number of ways:

- Numerically (descriptive statistics)
- Visually
- Statistically (inferential statistics)

In this course so far, we have done the first 2: by calculating the mean values of groups and by plotting histograms, density plots, and box-plots.

Today, we will be exploring the statistical side, using inferential statistics. Once again, we will be using our collars data, focusing on the battery life and the signal distance.

Let's get set-up by loading the tidyverse and reading in our data.

```
# Load library
library(tidyverse)

# Load data
collars <- read_csv("data/collar_data.csv")
```

Let's take a quick look at the data structure to remind ourselves what data we are using.

```
# View first few rows
head(collars)
```

```
# A tibble: 6 x 5
  collar_id maker          battery_life signal_distance  fail
      <dbl> <chr>                 <dbl>           <dbl> <dbl>
1         1 Collarium Inc.         141.           4171.     0
2         2 Collarium Inc.         121.           4134.     0
```

```
3         3 Collarium Inc.         126.          4277.     1
4         4 Collarium Inc.         127.          4198.     0
5         5 Collarium Inc.         141.          4173.     1
6         6 Collarium Inc.         105.          4175.     0
```

Before we dive in, if you need a refresher on inferential statistics, check out this powerpoint.

T-Tests

When we have a categorical variable with 2 categories, the statistical test that we use to determine if the two categories are *statistically significantly different* from each other is called a **t-test**.

When to Use a t-test

To run a t-test, the following things need to be true:

- The variable that we use to create our groups (***independent*** variable) is *categorical* and has only 2 categories.
- The variable that we want to compare between groups (***dependent*** or response variable) is *numeric*.

Independent vs. Dependent Variables

Why are the variables called *independent* and *dependent*, do you think?

- The **independent** variable is the *cause*. It does not change in response to other variables in the data.
- The **dependent** variable is the *effect*. We expect that it does change in response to the independent variable.

Want more info on independent vs. dependent variables? Check out this link.

Hypothesis Testing

The first step in running any statistical test (t-test or otherwise) is *hypothesis testing*.

We build the our hypotheses around our independent and dependent variables.

In this case, we want to know if there is a *meaningful* (read: statistical) difference between each collar manufacturer in their average values for battery life. To start, let's identify our independent and dependent variables:

- **Independent:** Maker, categorical
- **Dependent:** Battery_life, continuous/numerical

Based on our question, we can now set up two different statistical hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis always states that there is no difference or no relationship. We can think of the null hypothesis as our starting place—this is our default assumption. The alternative hypothesis is, well, the alternative to the null hypothesis. For example:

- **Null Hypothesis** (H_0): There is *no* difference in the means of battery life between the two collar makers
- **Alternative Hypothesis** (H_A): There *is* no difference in the means of battery life between the two collar makers

How can we determine whether we should *not reject* ("accept") or *reject* the null hypothesis? That's where statistics come in.

(Note: For reasons I won't go into, we never accept the alternative hypothesis, we only reject the null hypothesis.)

Running a t-test

There is a set of statistical tools that can help us whether or not there is a difference between 2 means. This group of tools are called t-tests.

Let's briefly remind ourselves of the logic here.

1. Our data are a *sample* of a larger *population* (think of the population as all of the collars ever produced by both companies).
    - Remember how we sampled our fish tank data for sick fish instead of getting data for every single tank? Same idea.
2. We're interested in the the difference in the means between the two groups.
    - If they're exactly the same, the difference in means would be 0.
    - If they're different, the difference between the means will be something either larger or smaller than 0.
3. However, because we only have data from a random *sample* of collars, there will be some variation in our numbers due to sampling error.
4. We want to know if the difference in means is due to sampling error alone or due to an actual difference between manufacturers.

The t-test allows us to determine if the means are different due to the random variation or if it represents an actual difference.

Let's run some code to perform our first t-test!

We use a function called `t.test()`.

- The first argument describes the test we want to run using column names. The structure is always `dependent ~ independent`.
- The second argument is the data frame we are referencing.

```
# A t-test for battery life
t.test(battery_life ~ maker, data = collars)
```

```
    Welch Two Sample t-test

data:  battery_life by maker
t = -15.966, df = 89.015, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Budget Collars LLC and group C
95 percent confidence interval:
 -38.98179 -30.35307
sample estimates:
mean in group Budget Collars LLC     mean in group Collarium Inc.
                        86.79449                         121.46192
```

Thankfully, the code isn't too onerous. Interpreting the results is a different matter, though…

Interpreting the Results of a t-test

Let's talk through it:

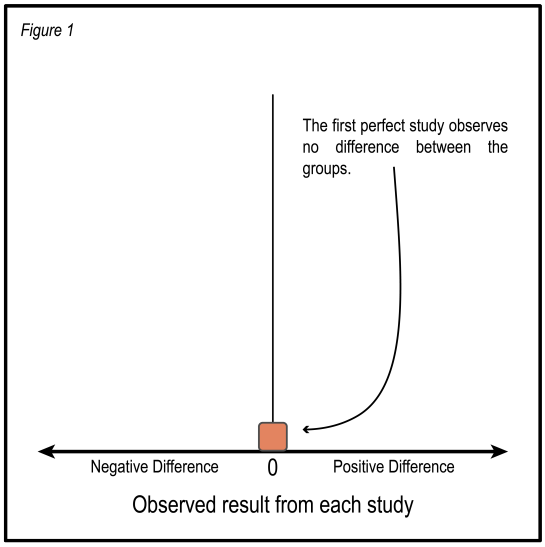
- `t`: this is what we call the t-value, or t-statistic.
  - Here, it is a metric of how different the difference of the two means is from 0.
  - Note that it doesn't correspond *directly* to the difference — it is taking into account the actual difference as well as the variation in the data sets.
  - Take home: big values (can be positive or negative) indicate a big difference from 0 while small values are mean the difference is close to 0.
- `df`: stands for *degrees of freedom*
  - It's a measure of your sample size and some other stuff — honestly, this is not something we're really going to focus on here.

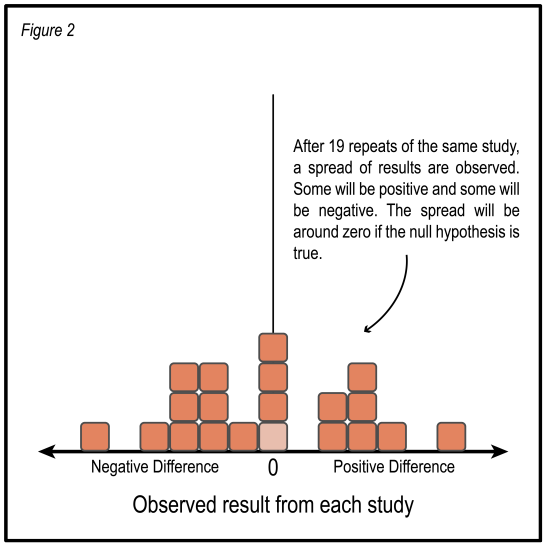
- `p-value`: a measure of certainty of the difference outlined in the `t-statistic` above
  - For the test above, our p-value is *very* small: 0.00000000000000022
    - Wait, how did I get from `2.2e-16` to 0.00000000000000022?
    - The `e-16` means that we need to move the decimal space to the left 16 times, creating a very small number!
    - If the number after the `e` were positive (e.g., `e+5`), I would move the decimal 5 places to the right, creating a very big number.
  - This means that there is an extremely low probability that the difference in means is due to random variation in our sample alone.
  - There are a lot of benchmarks in different fields for what this value should be below to consider the difference *significant*; typically a p-value below 0.05 is considered significant.
- `95 percent confidence interval`: this is the range that we can expect the test statistic (difference in means) to fall in 95% of the time given the data.
  - Our test statistic is the difference in means of Budget Collars and Collarium, or 86.8 - 121.5 = -34.7
  - If we randomly sample a group of collars from the "population", the difference in means will fall between -30.35 and -38.98 about 95% of the time

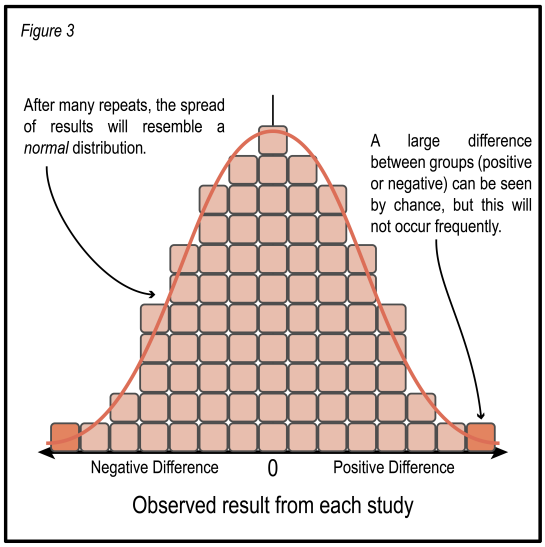
Let's quickly remind ourselves how p-values work:

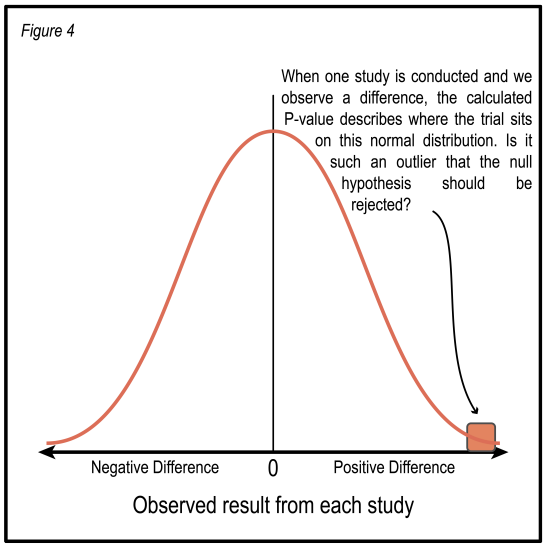
If the null hypothesis is true, what would be observed if we could repeat the study many times?

THE BOTTOM LINE









Given our very small p-value here, what can we conclude?

- Is the p-value below our cut-off for significance (0.05)?
- Is our result statistically significant?
- Does that mean we should or should not reject the null hypothesis?
- So...is there a "real" difference in the means between the two companies or not?

Let's Practice!

We want to know if there is a significant difference in the average signal distances for the two companies.

1. Identify your independent and dependent variables.

- **Independent:** maker, categorical
- **Dependent:** signal_distance, continuous

2. Write out your null and alternative hypotheses.

- **Null Hypothesis** (H_0): there is *no* difference in average signal distance between collar makers
- **Alternative Hypothesis** (H_A): there *is* a difference in average signal distance between collar makers

3. Run the t-test

```
# A t-test for signal distance
t.test(signal_distance ~ maker, data = collars)
```

```
    Welch Two Sample t-test

data:  signal_distance by maker
t = 15.837, df = 92.104, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Budget Collars LLC and group C
95 percent confidence interval:
  93.80549 120.70737
sample estimates:
mean in group Budget Collars LLC     mean in group Collarium Inc.
                        4302.074                          4194.817
```

4. Interpret the t-test

- Is the p-value below our cut-off for significance (0.05)? ***Yes, $p < 0.05$***
- Is our result statistically significant? ***Yes!***
- Does that mean we should or should not reject the null hypothesis? ***REJECT***
- So… is there a "real" difference in the means between the two companies or not? ***Yes, there is a real difference***