Week 6 Day 1: Confidence Interval for Snowfall between Weather Patterns

Learning outcomes

- Use bootstrapping to find a confidence interval for a difference in means.
- Interpret a confidence interval for a difference in means.
- Use a confidence interval to determine the conclusion of a hypothesis test.

Review from last week

Last week we used cards to simulate what differences in average snowfall we might have seen if the null hypothesis was true. However, today we're not interested in deciding if we believe the means of these two groups are similar or different.

Today, we are interested in estimating what range of values the **true** difference in means might take on.

1. Based on the p-value you obtained from Thursday's activity, do you believe 0 is a plausible value for $\mu_{\text{El Nino}} - \mu_{\text{La Nina}}$?

Confidence interval

A **confidence interval** represents a range of plausible values for a population parameter. In this case, our population parameter is $\mu_{\text{El Nino}} - \mu_{\text{La Nina}}$, or the true difference in mean snowfall between El Nino and La Nina years.

These best way to estimate what range of values a parameter might have is to go out and collect more samples. However, that is often not feasible. So, instead we mimic this process by *resampling with replacement* from our original sample. This process is called **bootstrapping**.

Bootstrapping snowfall & weather patterns

When bootstrapping with two groups, we're assuming that the sample within each group is *representative* of other possible values in the population. Here, we are assuming that the years included in our sample are representative of the snowfall for other El Nino / La Nina years.

Because we **are not** assuming the null is true (that there is no difference in the means of these two groups), we **do not** combine the groups together. Rather, we keep the groups separate and sample from each group separately.

2. Let's walk through how we would carry out this process:

Step 1:

Step 2:

Step 3:

Step 4:

3. What statistic do we have after step 4?

4. Once we create a bootstrap distribution of 1000 simulations, at what value do you expect the distribution to be centered? Explain your reasoning.

Creating a bootstrap distribution in R

We will use the **infer** package (again) to make our bootstrap distribution. The process we used for this situation will look very similar to before, since all we are changing is the statistic we calculate!

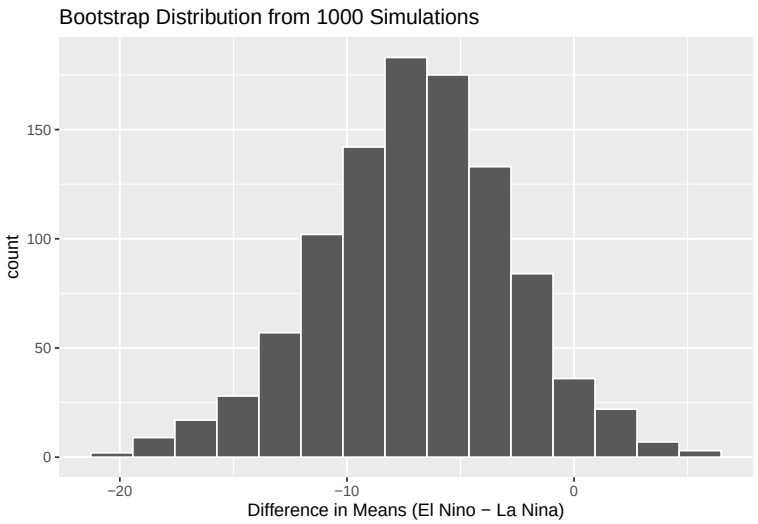
5. Fill in the blanks for the code below.

```
snow %>%

  specify(response = _________________, explanatory = _________________) %>%

  generate(reps = _________________, type = _________________) %>%

  calculate(stat = "diff in means",
            order = c("El_Nino", "La_Nina")
            )
```

6. What is the difference between this code and the code to generate a null distribution (what we did on Thursday)?

Obtaining a confidence interval

A bootstrap distribution from 1000 reps is plotted below.



6. What are the two methods we could use this distribution to obtain a confidence interval?

Percentile method

I've provided a table of different percentiles to help you create your confidence interval.

Quantile	Value
0.5%	-14.046067
1%	-17.594436
2.5%	-15.996228
5%	-14.046067
90%	-1.593442
95%	-0.234265
97.5%	1.265284
99.5%	-0.234265

7. Suppose we are interested in constructing a 95% confidence interval. Using the table above, report the end points of this confidence interval.

8. Interpret the confidence interval in the context of this investigation.

SE method

```
##        min        Q1    median        Q3      max      mean       sd    n
##  -20.25528 -9.701263 -6.889375 -4.198852 5.649167 -6.971641 4.235441 1000
##  missing
##        0
```

Using the t-distribution to create a confidence interval

Previously, we found our confidence interval by finding different percentiles on our bootstrap distribution. For example, we used the 2.5th and 97.5th percentile to obtain a 95% confidence interval.

When we are using a t-distribution to obtain our confidence interval, the process has similar ideas, but a slightly different approach. Since the t-distribution is centered at 0 and symmetric, the number associated with the 2.5th percentile and the 97.5th percentile **is the same**. Well, one is positive and one is negative, but they have the same numbers. So, we only need to find **one** number to make our confidence interval!

The number we are finding is called the **multiplier**. The multiplier for a confidence interval depends on two things, (1) the degrees of freedom and (2) the side of confidence interval you want. In our case we know we should use a t-distribution with 49 degrees of freedom.

6. We are interested in making a 95% confidence interval. Using the table below, circle the correct multiplier we should use to make our interval.

`R` code	Value
`qt(0.90, df = 49)`	1.299069
`qt(0.95, df = 49)`	1.676551
`qt(0.975, df = 49)`	2.009575
`qt(0.995, df = 49)`	2.679952

Now that we have the multiplier, we can put all of the pieces together! The "formula" for a t-based confidence interval is:

$$\text{point estimate} \pm t^*_{df} \times SE$$

Interpret the interval you calculated in question 19.

21. Would the results from a theory-based test match the results we saw with the simulation? Explain why or why not.

Take-home messages

1. To create one simulated sample on the bootstrap distribution for a difference in sample means, label $n_1 + n_2$ cards with the original response values. Keep groups separate and randomly draw with replacement n_1 times from group 1 and n_2 times from group 2. Calculate and plot the resampled difference in means.