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!

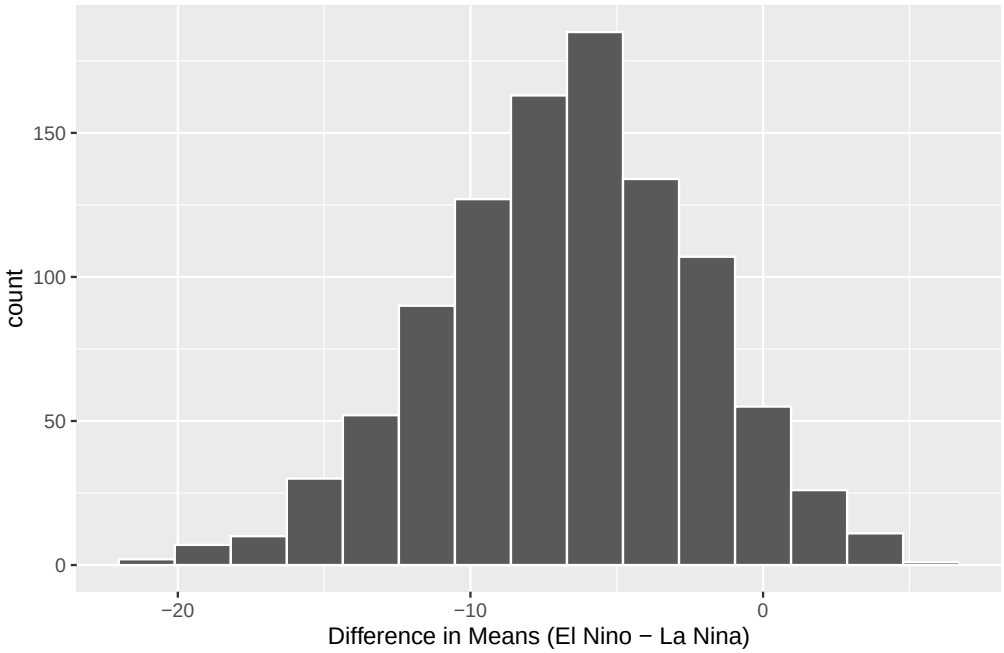
5. Fill in the blanks for the code below.

```
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.



7. What are the two ways 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%	-14.2069865
1%	-17.7884868
2.5%	-15.3915909
5%	-14.2069865
90%	-1.1118634
95%	0.5023083
97.5%	1.5679230
99.5%	0.5023083

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

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

SE method

A percentile confidence interval uses **only** the bootstrap distribution. The SE method on the other hand uses information from both the bootstrap distribution and the t-distribution.

Because this method uses a t-distribution it should only be used **if the bootstrap distribution is bell-shaped and symmetric**.

10. Do you believe this condition is violated?

Alright, let's see how this confidence interval works. Our formula looks like this:

$$\bar{x}_{\text{El Nino}} - \bar{x}_{\text{La Nina}} \pm t^*_{df} \times SE_{boot}$$

There are three pieces to the interval:

- the observed statistic ($\bar{x}_{\text{El Nino}} - \bar{x}_{\text{La Nina}}$)
- the t-distribution multiplier (t^*_{df})
- the standard error from the bootstrap distribution (SE_{boot})

11. What is the observed statistic for this investigation?

12. Using the table below, what is the standard deviation for the bootstrap distribution (the estimated standard error)?

```
##        min        Q1    median        Q3      max      mean       sd    n
##  -21.87756 -9.610177 -6.439658 -3.790266 4.927126 -6.666427 4.399483 1000
##  missing
##        0
```

13. Using the table below, circle the correct multiplier we should use to make our interval.

R code	Value
`qt(0.90, df = 20)`	1.3253407
`qt(0.90, df = 22)`	1.3212367
`qt(0.90, df = 43)`	1.3015516
`qt(0.95, df = 20)`	1.7247182
`qt(0.95, df = 22)`	1.7171444
`qt(0.95, df = 43)`	1.6810707
`qt(0.975, df = 20)`	2.0859634
`qt(0.975, df = 22)`	2.0738731
`qt(0.975, df = 43)`	2.0166922
`qt(0.995, df = 20)`	2.8453397
`qt(0.995, df = 22)`	2.8187561
`qt(0.995, df = 43)`	2.6951021

14. Using your answers to questions 11, 12, and 13, create a 95% confidence interval for the difference in mean snowfall between El Nino and La Nina years.

15. What value do we hope is contained in this interval?

16. Do we know if our interval contains this value?

Using the t-distribution to create a confidence interval

So far we've found a confidence interval using the percentile and SE methods. Both of these used some aspect of the bootstrap distribution. One final option is to use the t-distribution to create our confidence interval.

17. What distribution does a bootstrap distribution approximate?

18. If we wanted to use a t-distribution to approximate this distribution, what conditions do we need to check?

When we use theory-based methods to obtain our confidence interval, we use formulas to approximate the true standard error of the sampling distribution. So, where we used the standard deviation of the bootstrap distribution, now we will use a mathematical formula.

The formula for calculating the standard error of $\bar{x}_{\text{El Nino}} - \bar{x}_{\text{La Nina}}$ is:

$$SE = \sqrt{\frac{s^2_{\text{El Nino}}}{n_{\text{El Nino}}} + \frac{s^2_{\text{La Nina}}}{n_{\text{La Nina}}}}$$

19. Using the formula above, calculate the estimated standard error of the sampling distribution.

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

$$\overline{x}_{\text{El Nino}} - \overline{x}_{\text{La Nina}} \pm t^*_{df} \times SE$$

20. Using the multiplier you found in #13, calculate a 95% confidence interval for $\mu_{\text{El Nino}} - \mu_{\text{La Nina}}$.

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.
2. When using a bootstrap distribution to obtain a confidence interval, there are two methods you can use: the percentile method and the SE method.
3. The SE method for creating confidence intervals requires the bootstrap distribution be bell-shaped and symmetric.
4. The percentile method makes no assumptions about the shape of the bootstrap distribution.
5. You can choose between simulation-based methods (e.g., bootstrapping) and theory-based methods (e.g., t-distribution) to create a confidence interval.
6. Simulation-based conditions only require that the observations are independent.
7. Theory-based methods require that the observations are independent **and** that the distribution of each group is nearly normal.
8. If the conditions for theory-based methods **are not** violated, then both methods (theory & simulation) will yield similar results.
9. If the conditions for theory-based methods **are** violated, then the methods **will not** yield similar results.
10. If the conditions for theory-based methods **are** violated, using a t-distribution to find a p-value will **underestimate** the true p-value. A t-distribution will also result in a confidence interval that is **too narrow**!