Week 5 Day 4: Weather Patterns and Record Snowfall

Learning outcomes

- Given a research question involving one categorical explanatory variable and one quantitative response variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a simulation-based hypothesis test for a difference in means.
- Interpret and evaluate a p-value for a simulation-based hypothesis test for a difference in means.

Weather patterns and record snowfall

In the winter of 2018–2019, Bozeman, Montana had a record snowfall which resulted in the collapse of two flat-roofed buildings on the MSU campus. A writer for the *Washington Post* predicted the heavy snowfall for 2018–2019 due to the El Nino weather pattern that occurred in that season. A meteorologist in Montana wanted to see if the weather pattern really was associated with total snowfall. She obtained historical data from 44 years on the weather pattern (El Nino or La Nina and snowfall (in inches) at the Billings Weather Station.

Here's a preview of what the data look like:

```
glimpse(snow)
```

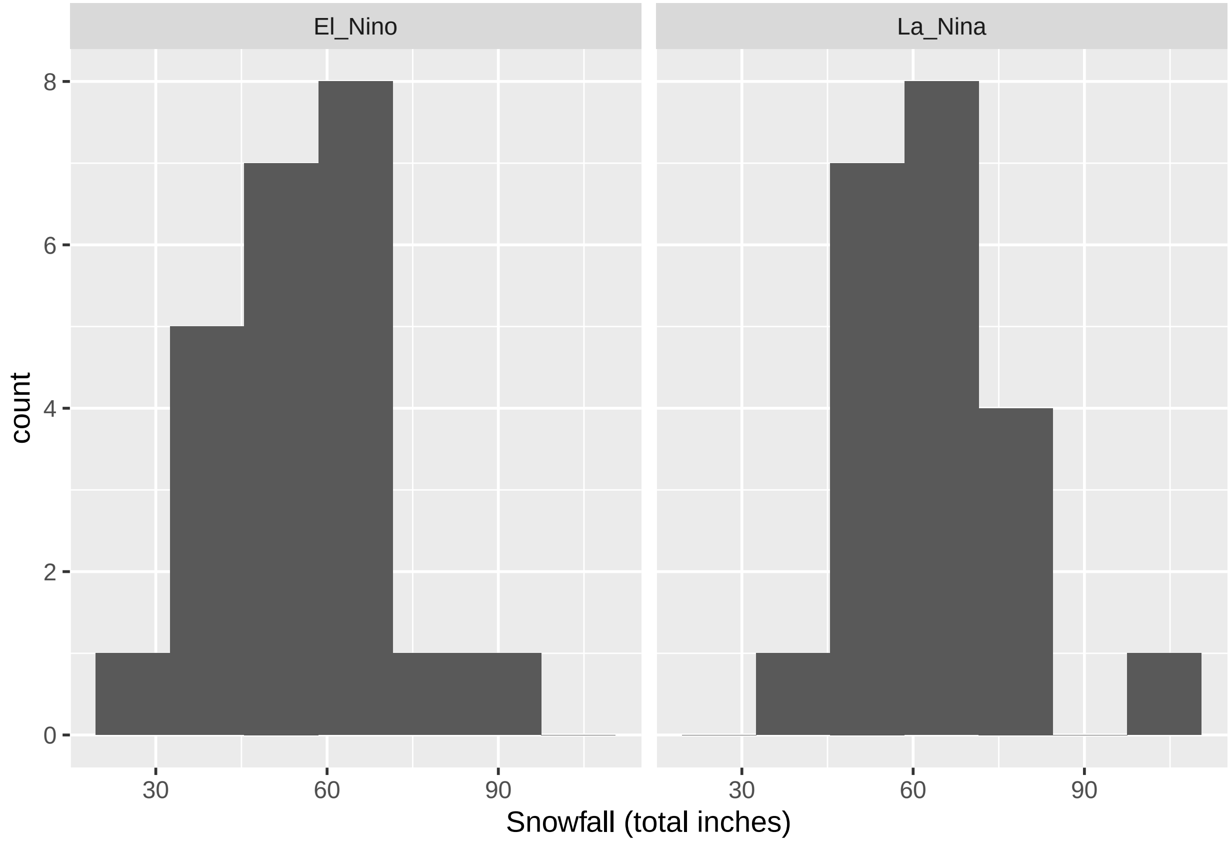
```
## Rows: 44
## Columns: 2
## $ WeatherPattern <chr> "El_Nino", "El_Nino", "La_Nina", "La_Nina", "El_Nino", ~
## $ Snowfall       <dbl> 31.9, 57.9, 107.2, 82.9, 65.8, 66.2, 60.1, 55.8, 42.8, ~
```

1. What is the observational unit for this study?

Exploratory Data Analysis

Let's start off by making some faceted histograms and calculating some summary statistics for each weather group.

```
ggplot(data = snow,
       mapping = aes(x = Snowfall)) +
  geom_histogram(binwidth = 13) +
  labs(x = "Snowfall (total inches)") +
  facet_wrap(~WeatherPattern)
```



```
favstats(Snowfall ~ WeatherPattern, data = snow)
```

```
##   WeatherPattern  min   Q1 median   Q3   max     mean       sd  n missing
## 1         El_Nino 31.9 46.4   57.7 64.3  87.9 56.23043 13.00823 23       0
## 2         La_Nina 44.5 51.4   60.9 70.3 107.2 63.13333 15.48626 21       0
```

2. The two variables assessed in this study are the type of weather pattern and snowfall. Identify the role for each variable (explanatory or response).

Explanatory:

Response:

3. Which group (El Nino or La Nina) has the highest center in the distributions of snowfall? Explain which measure of center you are using!

4. Using the faceted histograms, which group has the largest spread in snowfall? How did you make that choice?

5. Is this an experiment or an observational study? Justify your answer.

6. Do you believe the observations in the data are independent? Explain your reasoning.

Use statistical inferential methods to draw inferences from the data

Step 1: Ask a research question

7. Write out the **parameter** of interest in context of the study. Use proper notation and be sure to define your subscripts. Use El Nino minus La Nina as the order of subtraction.

8. Write out the null hypothesis in words.

9. Write the alternative hypothesis in notation.

10. Calculate the observed statistic of interest (difference in means). Use El Nino minus La Nina as the order of subtraction. What is the appropriate notation for this statistic?

Step 2: Conduct a Hypothesis test

Remember that the null distribution is created based on the assumption the null hypothesis is true. In this study, the null hypothesis states that **there is no association / relationship between the two variables**. This means that the snowfall values observed in the data set would have been the same regardless of the weather pattern that year.

I've provided your group with a set of cards to use to simulate a sample that could have happened if the null was true.

11. How many cards will we start with?

12. What will we write on each card?

13. Next, we need to generate a dataset that could have happened if the null hypothesis was true. How do we do this?

14. Once we have generated our new dataset that could have happened if the null was true, what value do we calculate? *Hint*: What statistic are we calculating from the data?

15. Create one simulation using the cards provided. Is your simulated statistic closer to the null value of zero than the difference in means calculated from the sample? Explain why this makes sense.

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

Carrying out the simulation in R

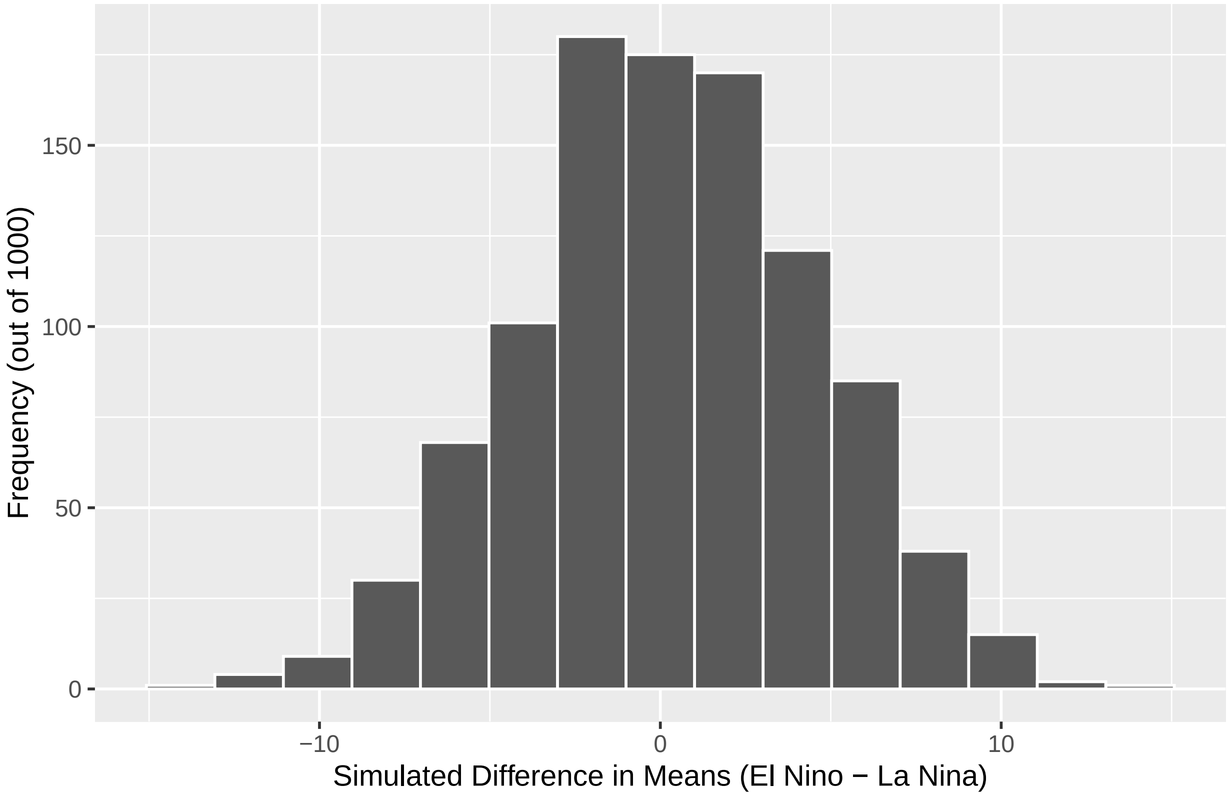
We will use the **infer** package (again) to make our simulated null distribution. The process we used for this situation will look very similar to before, since all we are changing is the statistic we calculate!

17. Fill in the blanks for the code below. You might want to look back at your Week 4 Day 2 activity (Inferences for Diving Penguins) for some help!

```
snow %>%

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

  hypothesise(null = _______________) %>%

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

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

Last time we use a "slope" statistic, so we didn't need to specify the order of subtraction. But now, with a difference in means we need to specify which group should come first and which should come second

18. Draw a line where the observed statistic falls on the simulated null distribution below. Shade the area that you will use to calculate the p-value.



19. Estimate the p-value for your hypothesis test. Based off of this p-value, write a conclusion to the hypothesis test.

Using theoretical methods instead...

What we just did used simulation to approximate what the sampling distribution of $\bar{x}_1 - \bar{x}_2$ would look like if the null was true. However, we don't necessarily need to use simulation to approximate this distribution!

The sampling distribution for $\bar{x}_1 - \bar{x}_2$ can be modeled using a t-distribution, when certain conditions are not violated. These conditions are:

- **Independence**: The sample's observations are independent
- **Normality**: Each sample should be approximately normal or have a large sample size. For *each* sample:
  - $n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a population whose distribution is nearly normal.
  - $n \geq 30$: If the sample size n is at least 30 and there are no particularly extreme outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

If these conditions seem reasonable, then we can use a t-distribution with the smaller of $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

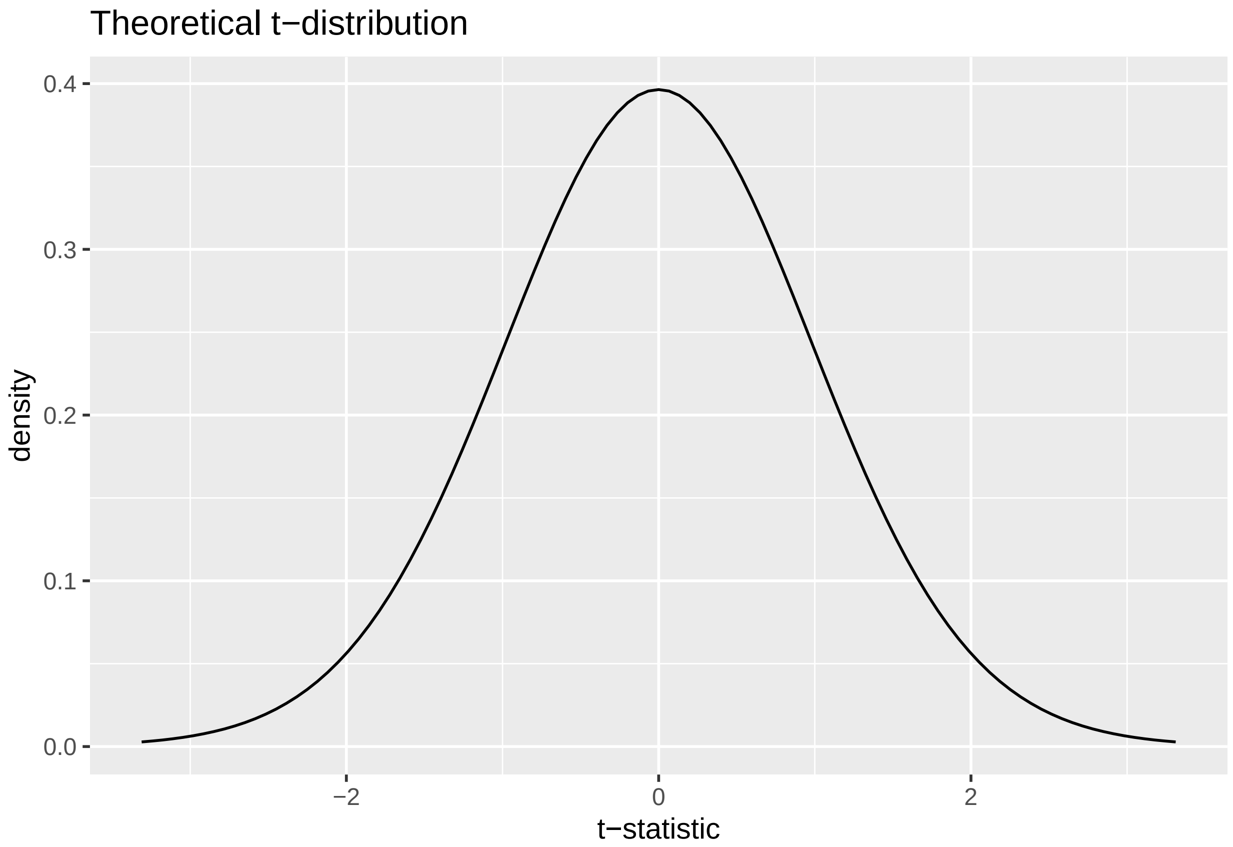
Observed Statistic

However, if we use a t-distribution, we need to use a **standardized statistic** (t-statistic) instead of the difference in means. To calculate a t-statistic we use the following formula:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

20. Using the above formula, calculate the t-statistic for these data.

21. Using the t-distribution below, find your calculated t-statistic. Shade the area that you will use to calculate the p-value.



22. Estimate the p-value for your hypothesis test. Is this similar to the p-value you obtained using simulation?

Take-home messages

1. To create one simulated sample on the null distribution for a difference in sample means, you carry out the following steps:

- label cards with the values from the observed sample
- tear the x and y values apart
- shuffle the cards and make new pairs of x and y values
- calculate and plot the difference in means

2. If it is not unreasonable to assume that the observations from each group come from a population with a normal distribution, then the t-distribution can be used (instead of a simulated null distribution) to approximate the sampling distribution.

- The t-distribution uses the smaller of $n_1 - 1$ and $n_2 - 1$ degrees of freedom