Inference for a Difference in Means

Learning Objectives.

- Write out the null and alternative hypothesis for One Categorical and One Quantitative Variable
- Calculate and carry-out theory-based hypothesis tests
- Interpret and evaluate a p-value
- Calculate and interpret a standardized statistic
- Construct and interpret a theory-based confidence interval
- Use a confidence interval to determine the conclusion of a hypothesis test

Background

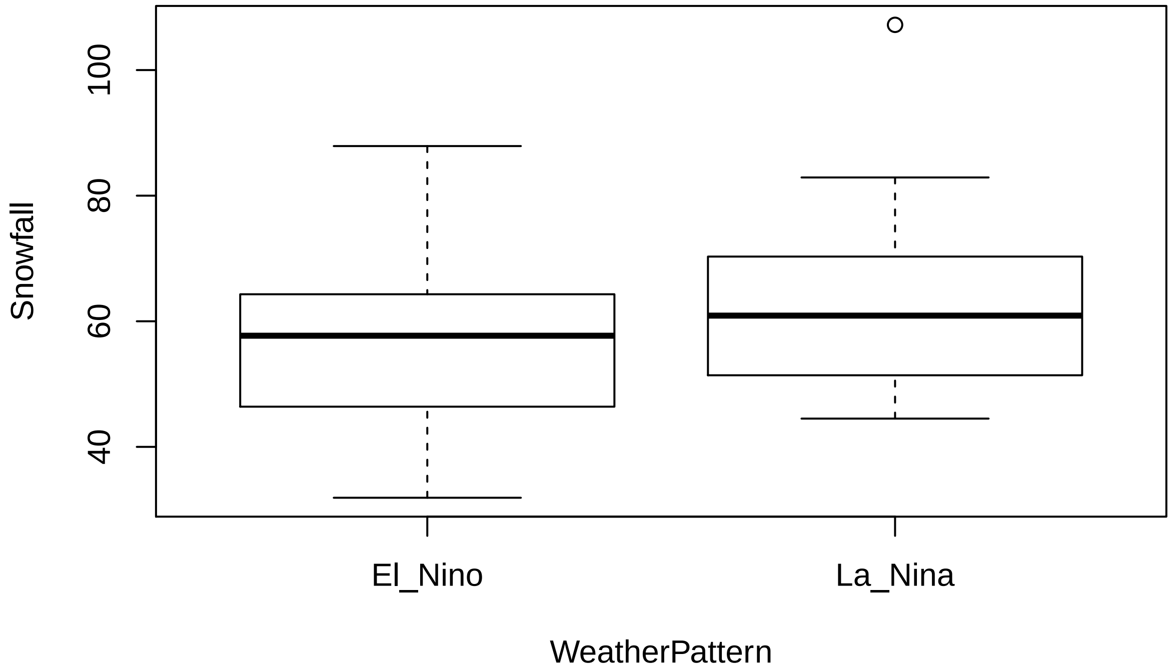
In the winter of 2018-2019, Bozeman had a record snowfall which resulted in the collapse of two flat-roofed buildings on the MSU campus. A writer for the Washington Post predicted the heavy snowfall for 2018-2019 due to the El Nino weather pattern that occurred in that season. A meteorologist in Montana 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.

```
# Set-up
library(readr)
library(car)
Snow <- read.csv("../data/SnowfallbyWeatherPattern.csv")
Snow$WeatherPattern <- factor(Snow$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
```

```
boxplot(Snowfall~WeatherPattern, data=Snow)
```



Quantitative Variables Review

1. The two variables assessed in this the type of weather pattern and snowfall. Identify the role for each variable (explanatory, response).

2. Which group (El Nino or La Nina) has the highest center? Explain what measure you are using?

3. Using the side-by-side boxplots, which group has the largest spread? How did you make that choice?

4. Is this an experiment or an observational study? Explain your reasoning.

Ask a research question.

5. Write out the parameter of interest in context of the study.

6. Write out the null hypothesis in notation.

7. Based on the research question, what is the direction of the alternative hypothesis?

Summarize and Visualize the data.

8. Calculate the summary statistic. Use El Nino minus La Nina as the order of subtraction. What is the appropriate notation for the statistic?

Use statistical inferential methods to draw inferences from the data

The T score measures the number of standard deviations the statistic is from the null value.

d$T = \frac{\bar{x_1} - \bar{x_2} - 0}{SE_{(\bar{x_1} - \bar{x_2})}}$, where $SE_{(\bar{x_1} - \bar{x_2})} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Remember to use the t distribution the following conditions must be met.

- The observations for each group must be normal for sample sizes less than For sample sizes greater than 60. . . .
- The observations in each group must be independent

9. Are these conditions met to use the t-distribution?

10. Calculate the point estimate.

11. Calculate the $SE_{(\bar{x_1}-\bar{x_2})}$.

12. Calculate the test statistic.

13. Interpret the value found in question 12.

To find the p-value we will use the function pt(test statistic, df, lower.tail=T). The test statistic is the value calculated in Q12, df is the smallest sample size minus 1. When you choose lower.tail = T this will give you the p-value for less than the test statistic. If you choose lower.tail = F this will give you the p-value for greater than the test statistic.

```r
pt(-1.59, df=20, lower.tail=T)
```

```
## [1] 0.06375917
```

15. How much evidence does the p-value provide against the null hypothesis?

Communicate the results and answer the research question.

To estimate the true difference in means we will use this equation for the confidence interval. $\bar{x}_d \pm t^* SE(\bar{x}_d)$

The t* multiplier is found in r using the following code:

```
qt(0.95+0.025, df=20)
```

```
## [1] 2.085963
```

16. Calculate the 95% confidence interval.

17. Interpret the interval you calculated in Question 16.