Paired Data

Learning Outcomes

- Given a research question, construct the null and alternative hypotheses in words and using appropriate statistical symbols
- Describe and perform simulation-based hypothesis for paired quantitative data
- Interpret and evaluate a p-value
- Find a confidence interval for the mean difference using bootstrapping
- Use a confidence interval to determine the conclusion of a hypothesis test

Terminology

The following terms will be covered in this activity.

- Mean difference
- Paired data
- Independent groups
- Shifted Null Distribution

For further explanation of these topics see Section 6.2 in the textbook.

COVID-19 and Air Pollution

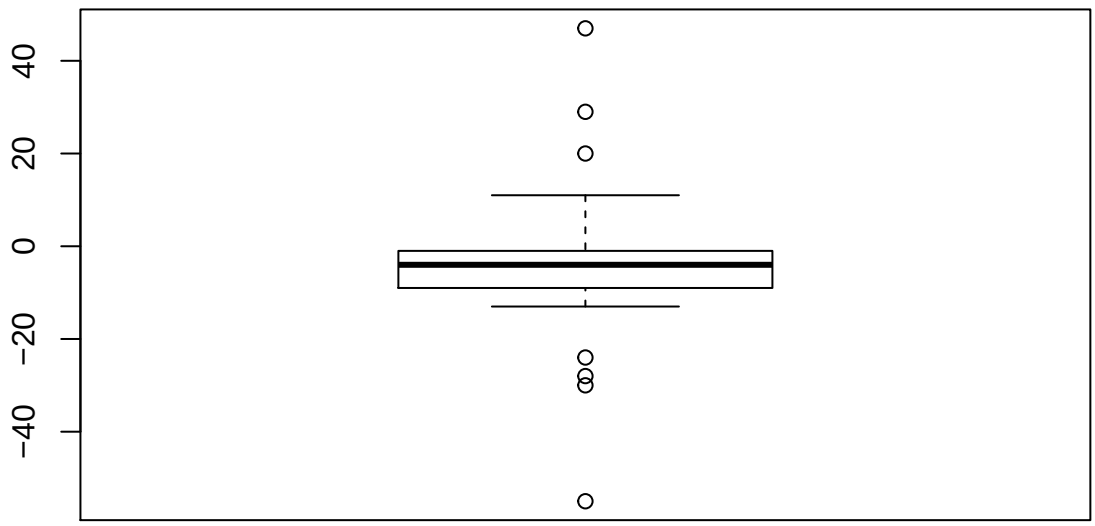
The social distancing efforts and stay-at-home directives to help combat the spread of COVID-19 have appeared to help 'flatten the curve' across the United States, albeit at a high cost to many individuals and businesses. The impact of these measures, though, goes far beyond the infection and death rates from the disease. You may have seen images comparing air quality in large international cities like Rome, Milan, Wuhan, and New Delhi such as the one pictured below which seem to indicate, perhaps unsurprisingly, that fewer people driving and factories being shut down have reduced air pollutants.

Have high population-density U.S. cities seen the same improved air quality conditions? To study this question, data was gathered from the U.S. Environmental Protection Agency (EPA) AirData website which records the ozone (O3) and fine particulate matter (PM2.5) values for cities across the U.S. These measures are used to calculate an air quality index (AQI) score for each city each day of the year. Thirty-three of the most densely populated U.S. cities were selected and the AQI score recorded for April 20, 2020 as well as the five-year median AQI score for April 20th (2015 - 2019). Note that higher AQI scores indicate worse air quality.

Vocabulary Review

```
Air <- read.csv("data/AirPollutionCOVID.csv")
```

```
boxplot(Air$Difference,
        main="Boxplot of the Differences in AQI Scores")
```



	Mean	Standard deviation	Sample Size
Current	$\bar{x}_1 = 47.394$	$s_1 = 14.107$	$n_1 = 33$
5 Year Median	$\bar{x}_2 = 51.545$	$s_2 = 17.447$	$n_2 = 33$
Differences	$\bar{x}_d$ = -4.152	$s_d = 17.096$	$n_d = 33$

1. What is the sample size?

2. Identify the variables in this study. What role do each have?

3. Why is this treated as a paired study design and not two independent samples?

4. Is this an experiment or observational study?

Ask a Research Question

5. What are the two competing possibilities to run a hypothesis test?

6. Write the null hypothesis in words.

7. What is the research question?

8. Write the alternative hypothesis in notation.

Summarize and Visualize the Data

9. Report the summary statistic for the data.

10. What notation is used for the value in question 9?

Use statistical inferential methods to draw inferences from the data

To simulate the null distribution we will use a bootstrapping method - sampling with replacement from the data set. Before bootstrapping we will need to shift the each data point by the difference $\mu_0 - \bar{x}$. This will ensure that the simulated null distribution will be centered at the null value.

11. Calculate the difference $\mu_0 - \bar{x}$. Will we need to shift the data up or down?

Add simulation here

12. Explain why the null distribution is centered at zero.

13. What proportion of samples are beyond the sample mean difference in AQI Scores for current scores minus 5 year median scores?

14. Interpret the p-value in the context of the problem.

15. How much evidence does this provide for improved air quality in US cities?

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

17. Use bootstrapping to find a 99% confidence interval for the parameter of interest. Fill in the .... Report the confidence interval in interval notation.

Communicate the results and answer the research question.

18. Interpret the 95% confidence interval in context of the problem.

19. Write a paragraph summarizes the results of this study. Be sure to include:

- Summary statistic
- P-value
- Interpretation of the p-value
- Confidence Interval
- Interpretation of the confidence interval
- Conclusion in context

Revisit and Look Forward