ACTIVITY 8

COVID-19 and Air Pollution

8.1 Learning outcomes

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

8.2 Terminology review

In today's activity, we will analyze paired quantitative data using simulation-based methods. Some terms covered in this activity are:

- Mean difference
- Paired data
- Independent groups
- Shifted null distribution

To review these concepts, see Section 6.2 in the textbook.

8.3 COVID-19 and air pollution

In June 2020, the social distancing efforts and stay-at-home directives to help combat the spread of COVID-19 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 in Figure @8.1, 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. A box plot of the differences in AQI scores for the 33 cities and a table of summary statistics are shown below.

Figure 8.1: The India Gate in New Delhi, India.

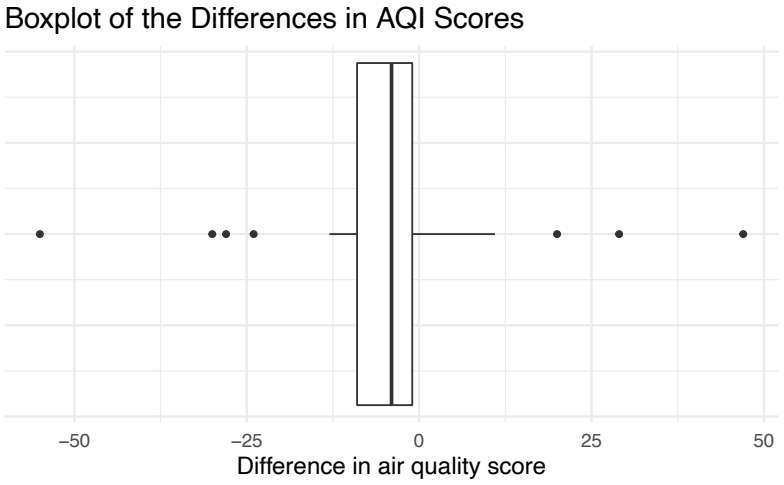


Table 8.1: Summary statistics for current AQI scores, median AQI scores from 2015–2019, and the differences in AQI scores.

	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	$s_d = 17.096$	$n_d = 33$

Vocabulary review

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? Justify your answer.

Ask a research question

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

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. Recall that the null distribution must be created under the assumption that the null hypothesis is true. Therefore, before bootstrapping we will need to shift each data point by the difference $\mu_0 - \bar{x}$. This will ensure that the mean of the shifted data is μ_0 and 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?

12. Use the provided `R` markdown file and enter the calculated value from question 11 for xx to simulate the null distribution and enter the summary statistic from question 9 for yy to find the p-value.

```
paired_test(data = Air$Difference,   #Vector of differences or data set with column for each group
        shift = xx,   #Shift needed for bootstrap hypothesis test
        as_extreme_as = yy,  #Observed statistic
        direction = "less",  #Direction of alternative
        number_repetitions = 1000,  #Number of simulated samples for null distribution
        which_first = 1)  #Not needed when using calculated differences
```

13. Sketch the null distribution created in question 12 here.

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

15. What proportion of samples are at or less than the sample mean difference in AQI Scores for current scores minus 5 year median scores?

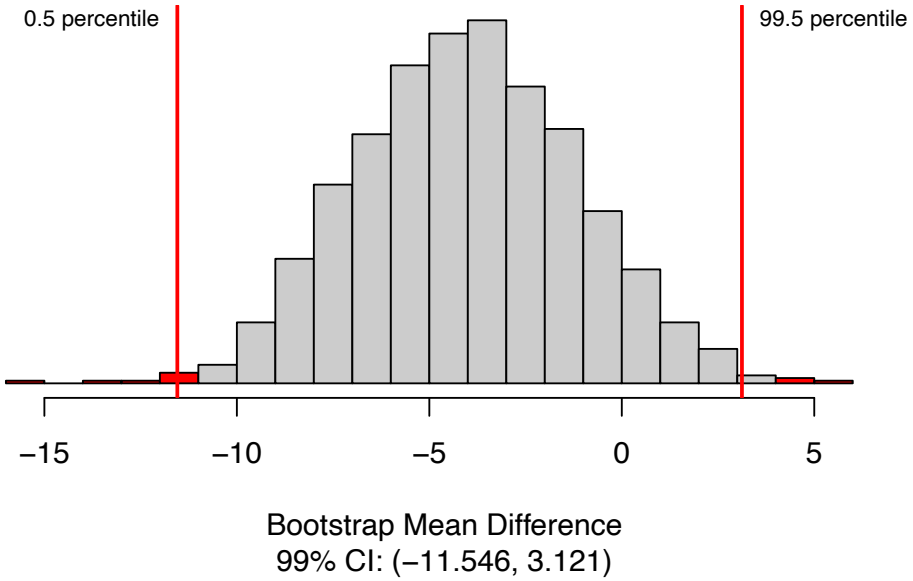
16. Interpret the p-value in the context of the problem.

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

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

The following R code creates a bootstrap distribution showing 1000 simulations of the mean difference.

```
paired_bootstrap_CI(data = Air$Difference, #Enter vector of differences
                    number_repetitions = 1000, #Number of bootstrap samples for CI
                    confidence_level = 0.99,  #Confidence level in decimal form
                    which_first = 1)  #Not needed when entering vector of differences
```



19. Use the bootstrapped distribution above to find a 99% confidence interval for the parameter of interest. Report the confidence interval in interval notation.

Communicate the results and answer the research question.

20. Interpret the 99% confidence interval in the context of the problem.

21. Do the results of your confidence interval and hypothesis test agree? What does each tell you about the null hypothesis?

22. Write a paragraph summarizes the results of this study. Be sure to describe:

- Summary statistic
- P-value and interpretation
- Conclusion (written to answer the research question)
- Confidence interval and interpretation
- Scope of inference

Revisit and look forward

23. Would it be possible to design an experiment to determine if the changed human behavior due to the COVID-19 pandemic causes a decrease in air pollution? Explain.

8.4 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity.