Week 6 Day 2: Paired Data, COVID-19 and Air Pollution

Learning outcomes

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

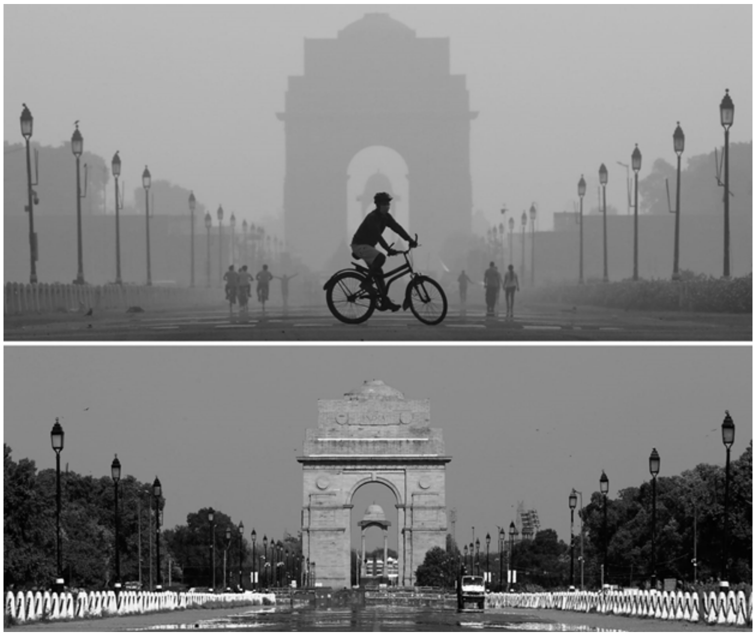


Figure 1: The India Gate in New Delhi, India.

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 the figures on the previous page, which seem to indicate, perhaps unsurprisingly, that fewer people driving and factories being shut down have reduced air pollutants.

Have high population-density US cities seen the same improved air quality conditions? To study this question, data were gathered from the US Environmental Protection Agency (EPA) AirData website which records the ozone (O3) and fine particulate matter (PM2.5) values for cities across the US. 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 US 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

1. Identify the variables in this study. What role (explanatory or response) do each have?

2. Are the **differences** in AQI scores independent for each case (US city)? Explain.

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

Ask a Research Question

4. What are the two competing possibilities for explaining the differences in AQI scores?

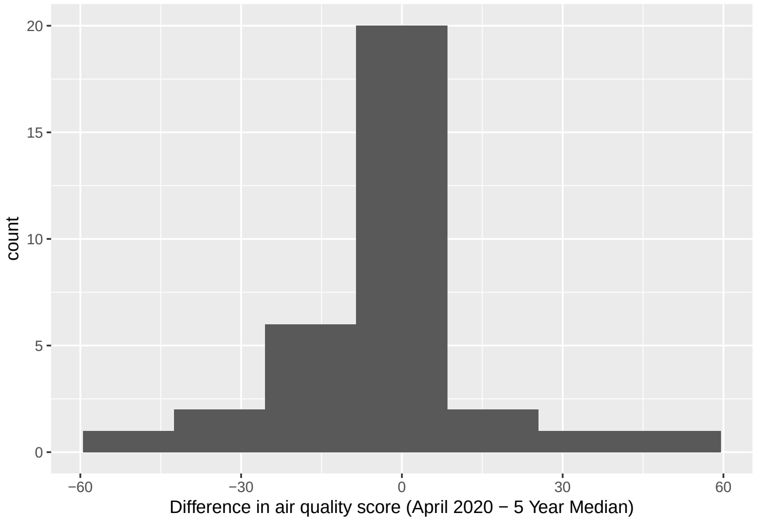
5. Write the null hypothesis in **words**.

6. What is the research question this study seeks to address?

7. Write the alternative hypothesis in **notation**.

Summarize and Visualize the Data

A histogram of the differences in AQI scores for the 33 cities and a table of summary statistics are shown below.



Summary statistics for the June 2020 (Current) AQI scores, median AQI scores from 2015–2019, and the differences in AQI scores.

Variable	Mean	Standard Deviation
Current	47.39	14.11
5yearMedian	51.55	17.45
Difference	-4.152	17.1

8. Report the summary statistic of interest (mean difference) for the data.

9. What notation is used for the value in question 8?

Use Statistical Methods to Draw Inferences from the Data

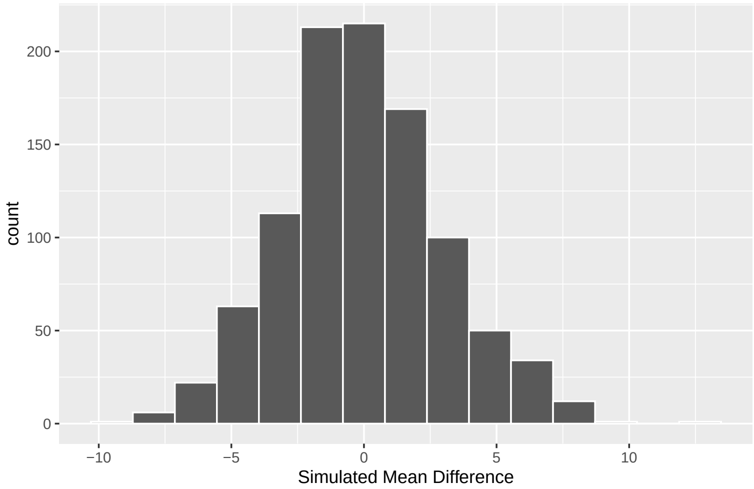
Hypothesis Test

To simulate the null distribution of paired sample mean differences we need to randomly assign the AQI score for each city to be either the current value or the five year median.

Take for example Albuquerque, New Mexico. In April 2020 the current AQI score was 23 and the five year median was 30. If we assume the null is true, that the mean of the differences is 0, then there is no relationship between Albuquerque's AQI score and the time period. That means, Albuquerque would have been just as likely to see an AQI score of 30 in April of 2020.

10. How can we use a coin to decide which values for each city are randomly assigned to the "Current" (April 2020) time period?

A simulated null distribution is shown below.



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

12. Find the observed mean difference on the distribution.

13. Shade the area of the distribution you will use to calculate the p-value.

14. Estimate the p-value for testing if there was improved air quality in US cities in April 2020.

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

16. If evidence was found for improved air quality in US cities, could we conclude that the stay-at-home directives ***caused*** the improvement in air quality? Explain.

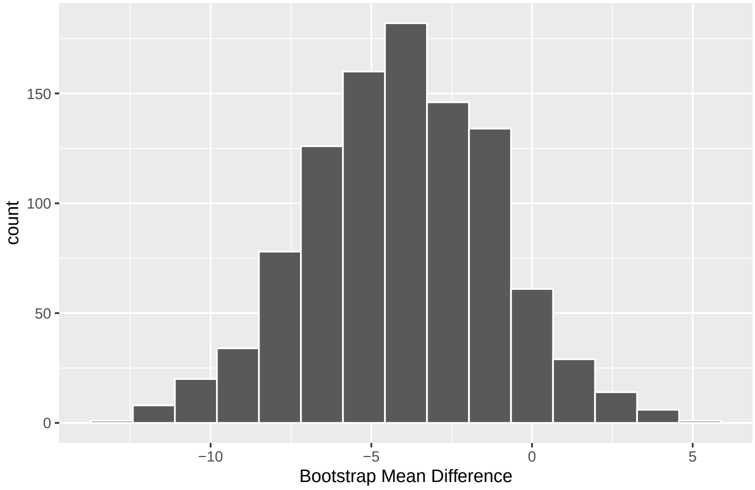
Confidence Interval

The goal of a confidence interval is to estimate a plausible range of values for the population parameter.

17. What is the population parameter in this study?

To create a bootstrap distribution, we **do not** assume the null hypothesis is true. Instead, we assume that the difference in AQI scores for these 33 US cities are representative of the differences in AQI scores for other US cities. So, we will randomly sample, with replacement, from our original sample to obtain a bootstrap sample.

I've created a bootstrap distribution below, using 1000 reps.



18. Where is the bootstrap distribution be centered? Why is it centered there?

19. Use the table below to find a 99% confidence interval for μ_{diff}.

Quantile	Value
0.5%	-8.848
1%	-10.91
2.5%	-9.91
5%	-8.848
90%	-0.3636
95%	0.6379
97.5%	1.67
99.5%	0.6379

Communicate the Results and Address 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?

Take-home messages

1. The differences in a paired data set are treated like a single quantitative variable when performing a statistical analysis. Paired data (or paired samples) occur when pairs of measurements are collected. We are only interested in the population (and sample) of **differences**, and not in the original data.

2. When analyzing paired data, the summary statistic is the "mean difference" **not** the "difference in means"[1]. This terminology will be *very* important in interpretations.

3. To create one simulated sample on the null distribution for the mean difference, we focus on each observation **not** on the groups of observations. For each observation, we flip a coin to decide which response value goes first and which goes second. We do this for **every** observation. Once we've randomly assigned which observation comes first, we find the difference in the values for each observation. Finally, we calculate and plot the simulated mean difference.

4. To create one simulated sample on the bootstrap distribution for a sample mean or mean difference, label n cards with the original values / differences. Randomly draw with replacement n times. Calculate and plot the resampled mean or mean difference.

[1] Technically, if we calculate the differences and then take the mean (mean difference), and we calculate the two means and then take the difference (difference in means), the value will be the same. However, the *sampling variability* of the two statistics will differ, as we will see in the next activity