Project EDDIE: Environmental Pollution & Public Health

Instructor Key

By the end of this module you will be able to answer the question: How does environmental pollution impact human health?

Learning Objectives

By the end of this module students will be able to:

* Calculate a cancer mortality rate
* Describe how environmental pollutants are correlated with cancer mortality rates
* Create a bar graph in Excel
* Execute a comparison of means t-test in Excel
* Determine if the means of two data sets are statistically distinguishable

Why This Matters

A healthy community begins with a healthy environment. Other health measures such as diet, exercise, and vaccination can all be undermined by polluted air and water. In conjunction with a strong public health system, an emphasis on environmental quality can reduce disease and increase well-being. Since the beginning of the industrial revolution, certain industries and practices have polluted the environment with carcinogens, endocrine disruptors, airway and skin irritants, and other toxicants. The presence of these pollutants has serious implications for human health, a conclusion supported by analysis of large population data sets, as will be demonstrated in this module.

Outline

Part A: Cancer Mortality Rates in the US

Part B: Louisiana’s Cancer Alley

Part C: Mountaintop Removal Coal Mines

Part A: Cancer Mortality Rates

The goal of public health practitioners is for every member of the community to live a long and healthy life. To help understand the underlying processes that impede this goal, public health researchers use a number of statistical metrics to aggregate health data.

One metric that scientists use to understand human health outcomes is mortality rate. Since population sizes are different (e.g. Alaska has a much smaller population than California), using total deaths from a specific cause is not a useful way to compare health outcomes, and scientists must normalize the data to account for variations in population size, age, and other demographics. In other words, a mortality rate makes data from populations of different sizes comparable, because it normalizes the number of deaths to the size of the population. For state-level data, mortality rates are calculated using the equation below. This calculation allows deaths to be expressed as deaths per 100,000 people. You can read more about calculating cancer mortality rates at the [National Cancer Institute](https://seer.cancer.gov/statistics/types/mortality.html).

$$Mortality Rate per 100,000=\frac{Number of Deaths}{Population} × 100,000$$

1) Consider the following data. This data set is the number of deaths from heart disease from the states listed and the approximate population of those states in 2019. Calculate the heart disease mortality rate for each of the states listed.

|  |  |  |  |
| --- | --- | --- | --- |
| State | Heart Disease Deaths in 2019 | Population in 2019 | Heart Disease Mortality Rate per 100,000 people |
| Alabama | 13,448 | 6,123,000 |  |
| Alaska | 843 | 649,000 |  |
| Arizona | 12,587 | 9,393,000 |  |
| Arkansas | 8,669 | 3,827,000 |  |
| California | 7,762 | 5,669,000 |  |

For the rest of this module we’ll be using cancer mortality rates collected from the [National Center for Health Statistics](https://www.cdc.gov/nchs/pressroom/sosmap/cancer_mortality/cancer.htm), organized by the CDC. You can download the data from the bottom of the linked page, or your instructor will provide it to you.

2) Open the data in Microsoft Excel. If you downloaded the data from the website you will need to save the data as an Excel workbook. If you were given the data by your instructor, it is already in this format. Look at the headers of the data in the tab titled ‘CancerRates.’ Fill in the table below with what column each of the data are stored in.

|  |  |
| --- | --- |
| Data | Column Letter |
| State |  |
| Year |  |
| Death Rate |  |
| Total Deaths  |  |
| Link to Data Source |  |

The goal of this section of the module is to calculate and visualize the average cancer mortality rate per 100,000 for each state across the years 2019, 2018, 2017, 2016, 2015, and 2014. The Excel sheet also contains data for 2005, but we are going to exclude this set of data since there is such a large time gap between it and when the rest of the data was collected. To calculate the average cancer mortality rate, you will need to reorganize the data.

3) Somewhere in the space to the right of the data, set up a table as shown below, with the left-most column containing the state names, followed by 6 columns of data, one for each year of data, and a column that will contain the average for each state. An example of how this can be set up is in the figure below.



4) Copy and paste the data into the appropriate columns.

5) At the bottom of the state column, add one more label called “National Average.” Use the Excel formula =average() to calculate the average of all the states for each year at the bottom of each column, as shown in the example below. After completing the first formula, you can copy and paste the formula to the right underneath the other columns.



6) In the column labeled “Average,” use the same Excel formula to calculate the average for each state across the 6 years of data, as shown in the example below. After completing the first formula, the equation can be copied and pasted down for all the states and the national average.



7) To make the data easier to interpret, the next step is to plot it as a bar graph. We want to plot just the state names as labels and the averages, which will be represented by the length of the bars. This means the state names and averages will need to be in columns next to each other, with the state names on the left. Copy and paste the state names again to a blank column over to the right. Then, copy and use paste special->values to paste the averages in the column to the right of the state names.

8) Highlight the state names and cancer mortality rate data. Then, under the “insert” menu, select the clustered column chart. An example of what this looks like on the iPad version of Excel is shown below. It might look slightly different in your version of Excel.



9) Use the “Elements” or “Add Chart Elements” (depending on your version of Excel) to add a y-axis label. Label the y-axis “Cancer Mortality Rate.” **You may need to widen the plot to see all of the labels on the x-axis. Be sure each column has a state label lining up with it. If it does not, make the graph longer.**

10) Looking at the plot, which state has the highest cancer mortality rate over the entire time period for which we have data?

Ans:

11) Name 2 other states that have cancer mortality rates higher than the national average.

Ans:

12) Which state has the lowest cancer mortality rate on average?

Ans:

13) Name 2 other states that have cancer mortality rates lower than the national average?

Ans:

Part B: Louisiana’s Cancer Alley

Now we will look at a case study from a single location to understand how environmental pollution impacts human health, using cancer mortality rates as a proxy for human health impacts.

Along the lower portion of the Mississippi River is an 85-mile stretch of land between Baton Rouge and New Orleans that houses almost 150 chemical factories and refineries. Located on the river, for ease of shipping, these factories and refineries are also located near, and in, some of the most populous areas of Louisiana. Residents of this area are exposed to emissions of hazardous toxicants at much higher concentrations than most of the United States population.

14) Look at the bar graph you made of cancer mortality rates in the United States. How does the cancer mortality rate of Louisiana compare to the national average?

Ans:

This area of Louisiana with all the refineries has not only a higher cancer mortality rate than the rest of the U.S., but also one of the highest rates in Louisiana. Due to the hotspots of cancer in this area, it’s earned the nickname “Cancer Alley.” However, just because the bar graph shows that one state has a higher or lower cancer rate than the national average doesn’t mean that the difference is meaningful or real. The spread of data between the highest and lowest points might be so big that the two sets of data mostly overlap, so much so, that they’re indistinguishable even though the means look different. In other words, the difference of means is not *statistically* *significant*.

In statistics, there are methods to determine if two sets of data are significantly different. One such test is the comparison of means t-test. A t-test creates a *p-value*, which is a probability*. If the p-value is below some threshold, usually 0.05, then the difference between the means of the two sets of data is said to be significant.*

Excel has a formula to calculate a t-test. The formula is =t.test(). Within the parentheses there are four arguments. The first two arguments are the two sets of data to be tested. The first set of data is highlighted, followed by a comma, then the second set of data is highlighted. The next two arguments tell Excel what type of t-test to execute. For this module, we will be using a two-tailed t-test (put a 2 in the formula in Excel), and two sample unequal variances (put a 3 in the formula in Excel. The formula in Excel should be typed as follows:

=t.test(first set of data, second set of data, 2,3)

15) Somewhere on your Excel sheet, create a table. Label the table “t-test results.” In the left-hand column, put a label for the t-test you are executing. The right column will have the p-values. The first test you will execute in this table is to compare the cancer mortality rate of Louisiana to the National Average. Arrange that information in your table as shown below:

|  |  |
| --- | --- |
| T-Test Results: |  |
| Comparison | p-value |
| Louisiana vs. National Average |  |
|  |  |

16) In the cell next to the “Louisiana vs. National Average” label and below the “p-value” label, type the formula for a t-test, highlighting the data for the years 2014-2019 for Louisiana and then the national average. The p-value will appear in the cell. What is the p-value of the t-test? *Excel may give the answer in scientific notation. If you are not familiar with scientific notation, you can increase the number of decimal places by formatting the cell, which will make it appear as a normal number. Depending on where you put this table, Excel may format the p-value as a date, which would look like 00/00/00. If this happens, change the format of the cell from ‘date’ to ‘number.’*

Ans:

17) Does this p-value indicate that there is a significant difference between the cancer mortality rate of Louisiana and the national average?

Ans:

18) Do you think that the pollution in Louisiana’s Cancer Alley influenced this result?

Ans:

19) Explain your answer to the previous question.

Ans:

Note that this finding would need to be followed with more research on how the specific pollutants influence cancer for the link to be confirmed.

Part C: Particulate Matter and Mountain Top Removal Coal Mines

Using the skills you learned in the previous sections (making bar graphs and performing t-tests), we will compare pollution at different locations with and without pollutant sources, and then evaluate the impact of that pollution on cancer mortality rates decades later.

One common type of industrial pollution is fine particulate matter. Fine particulate matter is considered an atmospheric pollutant of concern, as it can lodge deep in the lungs of those exposed to high concentrations of the pollutant, damaging the lungs. This damage is particularly harmful for people who have underlying health conditions such as asthma.

Because of the potential for health impacts, two types of particulate matter are monitored by the Environmental Protection Agency (EPA): PM10 and PM2.5. The number indicates the diameter of the particles, 10 $μ$m or smaller for PM10 and 2.5 $μ$m or smaller for PM2.5. The EPA sets national standards for acceptable concentrations of these air pollutants, below which concentrations are considered ‘safe.’ The [current national standard](https://www.epa.gov/air-trends/particulate-matter-pm25-trends) for PM2.5 is 12.0 $μ$g/m3.The EPA has air monitoring stations all over the US that periodically collect air samples to monitor particulate matter and other pollutants.

There are many sources of PM2.5 pollution, including roadway sources such as brake dust and exhaust, fire, including forest fires, industrial processes, mining, and construction. Concentrations can change quickly- both spatially and temporally- depending on local activities and wind circulation patterns.

20) Go to the EPA’s website for [Outdoor Air Quality Data](https://www.epa.gov/outdoor-air-quality-data/air-data-concentration-map). Using the drop-down menus

 a. Set the pollutant to PM2.5

 b. Set the Continental US or State field to Continental US (usually this is the default).

 c. Set the start date to 7-14-1999

d. Set the number of days to 10 (even if this box is set to 10, you have to click it for the “Plot Data” option to appear)

 e. Click “Plot Data”

 f. Scroll down to see the simulation

21) This map plots PM2.5 concentrations at EPA monitoring locations across the US over the 10-day period you selected. In which locations were there unhealthy (for sensitive groups) air concentrations of PM2.5 during that time period?

Ans:

22) Pick one of these locations. What do you think caused the PM2.5 concentration to be so high in that location on that date? You can use the sources listed above as a guide.

Ans:

23) Change the start date of the simulation to another time of your choice and play the simulation. What start date did you choose?

Ans:

24) What are some differences in the PM2.5 concentration data during this new time period compared to the previous time period?

Ans:

25) Based on the location of the site and the possible sources of PM2.5, what do you hypothesize caused these differences?

Ans:

One source of PM2.5 pollution in the Appalachian Mountains is mountaintop removal coal mines. In mountaintop removal coal mining, explosives are used to reduce the mountain side to debris that can be sifted to find coal. Read more about mountaintop removal coal mines and see the progression of a mine as viewed from space at the [NASA Earth Observatory](https://earthobservatory.nasa.gov/world-of-change/Hobet).

Mountaintop removal coal mining occurs in 4 states in the US: Kentucky, West Virginia, Tennessee, and Virginia. For the next part of this activity, you will examine mountaintop removal coal mine sites in West Virginia.

26) Go to this interactive [map of coal mines in West Virginia](http://www.wvgs.wvnet.edu/GIS/CBMP/all_mining.html). In the menu on the left, de-select “Underground Mining.” This will ensure only mountaintop removal sites are shown. What are 3 counties that have many mountaintop removal coal mines?

Ans:

26) You will now compare PM2.5 concentrations at a location in West Virginia to a control location in a mountainous Appalachian state that doesn’t have mountaintop removal mining. Click on the sheet in your Excel workbook titled “PM2.5.” This sheet contains PM2.5 data from 1999 for several locations in West Virginia and control sites. The data were collected from the [Outdoor Air Quality Data from the EPA](https://www.epa.gov/outdoor-air-quality-data/download-daily-data) site The time period overlaps with the simulation from the NASA Earth Observatory you previously watched. For each site, you can see the date the air sample was collected and the average PM2.5 concentration from that date. Create a table somewhere on your Excel sheet like the table below:

|  |  |  |
| --- | --- | --- |
| Site | Average Concentration | P-value: |
| West Virginia (site ID) |  |  |
| Control (site ID) |  |

27) This sheet contains 6 locations in West Virginia. Pick one of the locations in West Virginia to analyze. Update the table with the site ID. The site ID is the number listed under the county. Note there are multiple sites in the same counties. The site ID is important to distinguish these locations from within the same county.

 a. What is the site ID of the site you chose:

 b. What county is this site located in:

c: Look at the map of coal mines in West Virginia and find the county your site is located in on the map. Describe the abundance of mountain top removal coal mines in this county:

d. Calculate the average PM2.5 concentration using all of the data for your site, using the Excel formula =average() in the appropriate box in your table.

e. How does this average PM2.5 concentration compare to the national standard of 12.0 $μ$g/m3:

28) Pick a control site from the same sheet. The control sites are from mountainous areas in other Appalachian states that do not have mountain top removal coal mines. They are any site not located in WV. There are 4 control sites for you to choose from. Update the table with the site ID.

 a. What is the site ID of the site you chose:

 b. What county and state is the site located in:

c. Replace “Control” in your Excel table with the name of the state your site is located in. Update the site ID.

d. Calculate the average PM2.5 concentration using all of the data for your site, using the Excel formula =average() in the appropriate box in your table.

e. How does this average PM2.5 concentration compare to the national average:

29) Using the table you made in Excel, create a bar plot of the averages. How do the PM2.5 concentrations compare between your two sites?

Ans:

30) Conduct a t-test of the two sites as you did in Part B. Put the formula in the cell under the label ‘p-value’ in your table.

 a. What is the p-value of the test:

 b. Is there a significant difference between the PM2.5 concentrations at both sites:

 c. What do you think is driving this difference:

31) Next, compare your mountain top removal site to another mountain top removal site. Add another row to the table you made in #26. Label this “West Virginia (site ID).” In the parentheses, put the site ID for this location.

 a. What is the site ID of the site you chose:

b. Calculate the average PM2.5 concentration using all of the data for your site, using the Excel formula =average() in the appropriate box in your table.

c. Expand your bar graph to include this new average. How does the PM2.5 concentration at this site compare to the other West Virginia site you chose:

d. Conduct a t-test between the two West Virginia sites, adding the p-value to your table. What is the p-value:

e. Is there a significant difference between the PM2.5 concentrations at both West Virginia sites:

f. What does this mean in the real world for the PM2.5 concentrations in counties with mountain top removal coal mines:

31) Now you will explore how the air quality from 1999 affected the cancer mortality rates of people living in those areas decades later. Often times health effects of environmental pollutants are not immediate, but take years or decades to appear. Go back to the “Cancer Mortality” sheet. Find the bar graph you previously made.

a. How does the cancer mortality rate of West Virginia compare to that of the state your control site is in:

b. How does the cancer mortality rate of West Virginia compare to the national average:

c. Add another row to the table you made in step 15. Label this new row West Virginia vs. whatever control state you chose. In the table, in the appropriate cell, complete a t-test comparing the cancer mortality rates of West Virginia and the state of the control site you chose. What is the p-value of this t-test:

d. Is there a significant difference between cancer mortality rates for these two states:

e. Based on the other data you analyzed during this module and your understanding of the role of PM2.5 in disease, what do you think could be one factor driving this difference:

32) What is the benefit of a p-value in analyzing the cancer mortality rates and PM2.5 concentrations in this module beyond the bar graphs you created?

33) Based on the data in this module, how does environmental pollution impact human health?

Ans:

34) Do you think the analyses you have performed in this activity are sufficient to bring a lawsuit against the polluters on behalf of those harmed? Explain your reasoning. If you think more information is needed, what information would that be?

Ans: