**Activity Handout**

**Purpose:** In this activity, you will investigate data collected on lead levels in the water of Flint homes. You will compute percentiles, create a graph, and examine the impact of omitting data.

**Instructions**: Record your answers to the questions below on a separate sheet of paper or in a Word document on your computer. These answers will be necessary when writing a report at the end of the assignment.

**Step 1:** Background reading. Read the following news stories:

<http://michiganradio.org/post/video-how-dropping-two-flints-lead-test-numbers-changed-things-state>

<http://michiganradio.org/post/expert-says-michigan-officials-changed-flint-lead-report-avoid-federal-action>

**Step 2:** Answer the following questions about the news stories.

1. What is the “action level” for lead content in water?
2. What statistic is computed to be compared to the action level?
3. How large was the Michigan Department of Environmental Quality (MDEQ) sample size supposed to be?
4. Which homes should be measured?
5. Which lead measurements were omitted by the MDEQ, and why?

Full sample (*n* = 71) – The numbers crossed out are the ones omitted in the altered sample.  
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 9, 10, 10, 11, 13, 18, ~~20~~, 21, 22, 29, 42, 42, ~~104~~

**Step 3:** Calculate the 90th percentile for the Flint water data. Do this on the full (*n* = 71) and altered (*n* = 69) samples. Perform the calculations by hand. Two methods for calculating the 90th percentile will be used. Be sure to include the details of your calculations so readers of your report can check your work. **Note:** In the index formulas below, for the 90th percentile.

***First method:*** This method was illustrated in the video by an Eastern Michigan University math professor. The index formula is *np* + 0.5 and interpolation is used if the index is not a whole number. Make this calculation on the full (*n* = 71) and altered (*n* = 69) samples.

***Second method:*** The index formula is *(n* – 1)\**p* + 1, and interpolation is used if the index is not a whole number. This is the formula used by CODAP. Make this calculation on the full (*n* = 71) and altered (*n =* 69) samples. **Note:** The Excel function PERCENTILE.EXC uses yet another index formula, *p(n*+1*).*

**Step 4a:** Use your Step 3 results to determine if it matters which percentile formula is used—not in general, but for this specific application. (***Hint:*** What is the lead action level, and would the city of Flint have been affected by which formula was used?)

**Step 4b:** Use your calculations to explain the effect of omitting the two measurements in question. Write freely for at least one paragraph, taking the information from the news articles and connecting it to what you have heard about the Flint water crisis.

**Step 5a:** Create a dot plot.

* Type the data into CODAP or import data from the .csv file your teacher gives you.
* Click the *Graph* icon.
* Drag the variable *Lead* to the x-axis of the graph.

**Step 5b:** Add reference lines for the percentile and action limit.

* On the graph menu, click the *Measure* icon that looks like a ruler. Select *Plotted Value.*
* At the top of the graph, click on *value =*.
* Type in the command for calculating the 90th percentile: percentile(Lead, 0.9)  
  The percentile should now show up as a green line. If you hover over the line, you’ll see the value you calculated using the second method on the full sample.
* On the graph menu, click the *Measure* icon that looks like a ruler. Select *Movable Line.*
* Drag the line until it’s placed at the Action Level (15 ppb). ***Hint:*** If you use the resizing arrows to make your graph wider, it will be easier to place the line at exactly 15.
* On the graph menu, click the icon that looks like a camera to save your graph as a PNG file. Alternatively, you may wish to use a snipping tool to capture the graph and paste it into a Word file for your report.

**Step 5c**: Show what happens when you alter the sample.

* Click on the dot representing Lead = 104.
* On the graph menu, click the icon that looks like an eye and choose *Hide Selected Case.*
* Repeat for the dot representing Lead = 20 so this point is also hidden.
* Note that the green line has moved. CODAP automatically recalculated the 90th percentile on the altered sample.
* Use the camera icon (or snipping tool) to capture this altered graph for your report.
* If you wish to go back to the original graph, click the icon that looks like an eye and choose *Show All Cases.*

***Note:*** Alternate instructions for Step 5, describing how to create a detailed chart (including reference lines) in Excel, are included as a separate document.

**Step 6:** Write a report, using the results of your work in the previous steps. Include a brief summary of the readings in Step 1 and your answers to Step 2 in a short introduction. Include your Step 3 calculations, your answers to Step 4, the graphs from Step 5, and a conclusion.

**Activity Handout (with Correct Responses)**

**Purpose:** In this activity, you will investigate data collected on lead levels in the water of Flint homes. You will compute percentiles, create a graph, and examine the impact of omitting data.

**Step 1:** Background reading. Read the following news stories:

<http://michiganradio.org/post/video-how-dropping-two-flints-lead-test-numbers-changed-things-state>

<http://michiganradio.org/post/expert-says-michigan-officials-changed-flint-lead-report-avoid-federal-action>

**Step 2:** Answer the following questions about the news stories.

1. What is the “action level” for lead content in water? 15 parts per billion
2. What statistic is computed to be compared to the action level? The 90th percentile
3. How large was the Michigan Department of Environmental Quality (MDEQ) sample size supposed to be? 100
4. Which homes should be measured? The homes most at risk of contamination (homes known to either be serviced by lead service lines or that have lead pipes or pipes with lead solder in them)
5. Which lead measurements were omitted by the MDEQ, and why? 104 (because the house had a water filter), 20 (the sample came from a business)

Full sample (*n* = 71) – The numbers crossed out are the ones omitted in the altered sample  
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 9, 10, 10, 11, 13, 18, ~~20~~, 21, 22, 29, 42, 42, ~~104~~

**Step 3:** Calculate the 90th percentile for the Flint water data. Do this on the full (*n* = 71) and altered (*n* = 69) samples. Perform the calculations by hand. Two methods for calculating the 90th percentile will be used. Be sure to include the details of your calculations so readers of your report can check your work. ***Note*:** In the index formulas below, for the 90th percentile.

**First method:** This method was illustrated in the video by an Eastern Michigan University math professor. The index formula is *np* + 0.5 and interpolation is used if the index is not a whole number. Make this calculation on the full (*n* = 71) and altered (*n* = 69) samples.

The index is *np* + 0.5 = (71)(.90) + 0.5 = 64.4, so the 90th percentile is between the 64th (18) and

65th (20) ordered lead level values. We must interpolate to get 18.8 ppb for the 90th percentile. 18.8 is 40% of the way between 18 and 20. 2(0.4) = 0.8 and 18 + 0.8 = 18.8.

After we delete 104 and 20 from the data, *n* = 69, and the index is *np* + 0.5 = (69)(.90) + 0.5 = 62.6, so the 90th percentile is between the 62nd (11) and 63rd (13) ordered values. Once we interpolate, we get 12.2 ppb. 12.2 is 60% of the way between 11 and 13.

**Second method:** The index formula is *(n* – 1)\**p* + 1, and interpolation is used if the index is not a whole number. This is the formula used by CODAP. Make this calculation on the full (*n* = 71) and altered (*n =* 69) samples. ***Note*:** The Excel function PERCENTILE.EXC uses yet another index formula, *p(n*+1).

With all *n* = 71 values, *(n* – 1)\**p* + 1 = (71 – 1)\*0.9 + 1 = 64, so the 90th percentile is the 64th ordered lead level value, 18.

With only *n* = 69 values *(n* – 1)\**p* + 1 = (69 – 1)\*0.9 + 1 = 62.2 so the 90th percentile is between the 62nd (11) and 63rd (13) ordered lead values. We must interpolate to get 11.4.

**Step 4a:** Use your Step 3 results to determine if it matters which percentile formula is used—not in general, but for this specific application. (***Hint*:** What is the lead action level, and would the city of Flint have been affected by which formula was used?)

The choice of formula doesn’t affect the results very much. If they use the full sample, they will be above the action limit no matter which formula they use. If they use the altered sample, they will be below the action limit no matter which formula they use.

**Step 4b:** Use your calculations to explain the effect of omitting the two measurements in question. Write freely for at least one paragraph, taking the information from the news articles and connecting it to what you have heard about the Flint water crisis.

Omitting two measurements makes an important difference, because it changes the result from one that is above the action limit to one that is below the action limit …

**Step 5a:** Create a dot plot.

* Type the data into CODAP or import data from the .csv file your teacher gives you.
* Click the *Graph* icon.
* Drag the variable *Lead* to the x-axis of the graph.

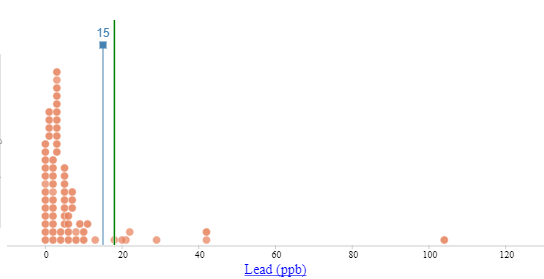
**Step 5b:** Add reference lines for the percentile and action limit.

* On the graph menu, click the *Measure* icon that looks like a ruler. Select *Plotted Value.*
* At the top of the graph, click on *value =*.
* Type in the command for calculating the 90th percentile: percentile(Lead, 0.9)  
  The percentile should now show up as a green line. If you hover over the line, you’ll see the value you calculated using the second method on the full sample.
* On the graph menu, click the *Measure* icon that looks like a ruler. Select *Movable Line.*
* Drag the line until it’s placed at the Action Level (15 ppb). ***Hint***: If you use the resizing arrows to make your graph wider, it will be easier to place the line at exactly 15.
* On the graph menu, click the icon that looks like a camera to save your graph as a PNG file. Alternatively, you may wish to use a snipping tool to capture the graph and paste it into a Word file for your report.

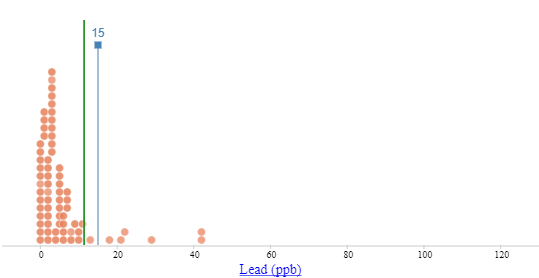
**Step 5c**: Show what happens when you alter the sample.

* Click on the dot representing Lead = 104.
* On the graph menu, click the icon that looks like an eye and choose *Hide Selected Case.*
* Repeat for the dot representing Lead = 20 so this point is also hidden.
* Note that the green line has moved. CODAP automatically recalculated the 90th percentile on the altered sample.
* Use the camera icon (or snipping tool) to capture this altered graph for your report.
* If you wish to go back to the original graph, click the icon that looks like an eye and choose *Show All Cases.*

Full sample – notice the 90th percentile (represented by the green line) is higher than the action limit.



Altered sample – notice the 90th percentile (represented by the green line) is lower than the action limit.



**Step 6:** Write a report using the results of your work in the previous steps. Include a brief summary of the readings in Step 1 and your answers to Step 2 in a short introduction. Include your Step 3 calculations, your answers to Step 4, the graphs from Step 5, and a conclusion.

Introduction: The Flint water crisis is a tragic story of a government blunder resulting in unsafe water piped into more than 100,000 homes. One aspect of the story involves botched data collection, which was made worse by the deletion of two measurements. Water was sampled from 71 addresses, and the lead content was determined. The EPA provides a rule for authorities to follow: If the 90th percentile of the lead values falls above 15 parts per billion (ppb), action must be taken to alert the people and correct the problem. The city of Flint was instructed to collect water samples from at least 100 homes. These homes are required to be most at risk of contamination, to give a “worst case scenario” estimate. Experts have criticized the selection of homes, saying they were not most at risk, and the sample size fell short. Here is where things get (more) suspicious. The Michigan Department of Environmental Quality (MDEQ) instructed the city of Flint to delete two lead values. The reason given was that they violated the testing guidelines of sampling homes most at risk. One address was a residence with a water filter installed, so the justification to delete it is that it was not at risk of having high lead values. In fact, it was the largest value by far at 104 ppb. The other deletion was found not to be a home; it was a business. This value was not as large, but was among the largest values at 20 ppb. As we will see, omitting these two values had a drastic effect on the estimation of the 90th percentile.

Calculations:

Here is the data with all *n* = 71 lead values:

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 9, 10, 10, 11, 13, 18, 20, 21, 22, 29, 42, 42, 104

There are several ways to compute a percentile. Two ways will be shown here.

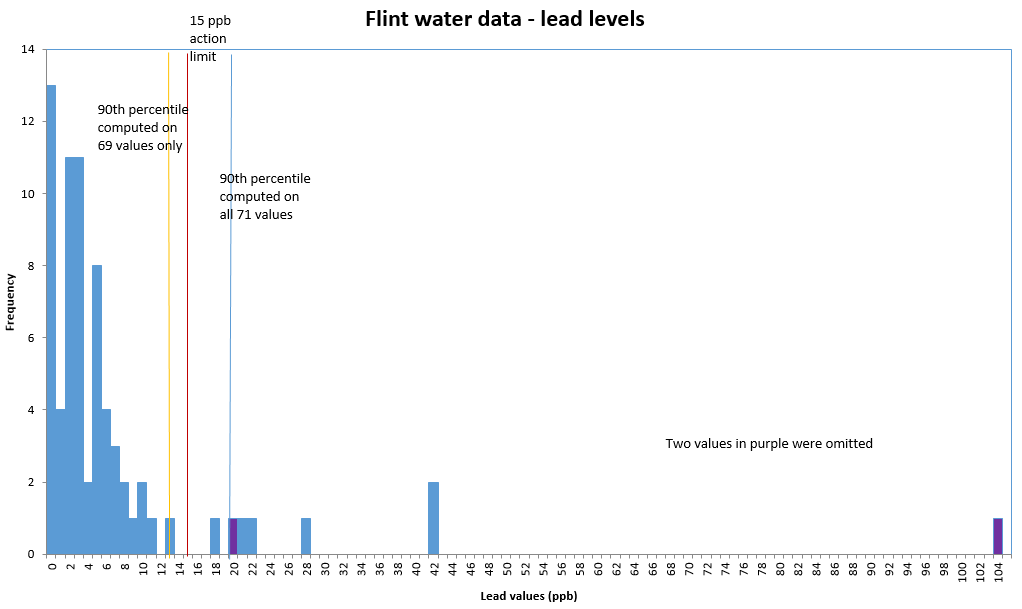
The first was demonstrated by a math professor in one of the background news stories we read. It uses *index =* *np* + 0.5. The index is *np* + 0.5 = (71)(.90) + 0.5 = 64.4, so the 90th percentile is between the 64th (18) and 65th (20) ordered lead level values. We must interpolate to get 18.8 ppb for the 90th percentile.

After we delete 104 and 20 from the data, *n* = 69 and the index is *np* + 0.5 = (69)(.90) + 0.5 = 62.6, so the 90th percentile is between the 62nd (11) and 63rd (13) ordered values. Once we interpolate, we get 12.2 ppb.

The second method uses *index* = (*n* – 1)\**p* + 1) and will be checked with CODAP later. With all *n* = 71 values, (*n* – 1)\**p* + 1 = (71 – 1)\*0.9 + 1 = 64, so the 90th percentile is the 64th ordered lead level value, 18. With only *n* = 69 values (*n* – 1)\**p* + 1 = (69 – 1)\*0.9 + 1 =62.2, so the 90th percentile is between the 62nd (11) and 63rd (13) ordered lead values. We must interpolate to get 11.4. The CODAP percentile function confirms the 2nd formula calculations.

The 90th percentile must be compared to the action level of 15 ppb. On the full *n* = 71 sample, it doesn’t matter which percentile formula is used. Either way, the computed 90th percentile, 18.8 or 18, falls above the 15 ppb action level. Once the two values are omitted, it still doesn’t matter which percentile formula is used. Either way, the computed 90th percentile, 12.2 or 11.4, falls below the 15 ppb action level.

It is good to know the index formulas don’t affect the results very much, but it is clear that deleting the two lead values had a huge effect on the results. Since the MDEQ didn’t have a good justification for deleting these values, it seems suspicious. Deleting those two values allowed the authorities to conclude the water was safe, since the 90th percentile of lead values fell safely under the 15 ppb action level.



The above graph shows the 71 lead values, with the two deleted points in purple. The vertical lines show the 90th percentile was over the 15 ppb action limit until the two values were deleted.

To create the graph shown above, use the directions provided in a [separate Excel file.](https://www.statisticsteacher.org/files/2019/09/Flint-dataset.csv)