UC Santa Barbara
GIS Core Curriculum for Technical Programs (1997-1999)

Title
Unit 39: Performing Statistical Analyses

Permalink
https://escholarship.org/uc/item/4p56m3j1

Authors
Unit 39, CCTP
Funk, Chris

Publication Date
1998

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

The NCGIA GIS Core Curriculum for Technical Programs

UNIT 39: PERFORMING STATISTICAL ANALYSES

Written by Chris Funk, Geography Department, University of California Santa Barbara

Context

People use GIS to answer questions about the world. Typically, these answers involve different types of generalizations. These generalizations may relate either to questions about the physical or human reality around us. For example:

- What is the average temperature of Botswana in June? Of Kuala Lumpur in August?
- What is the average distance that people are willing to drive to the grocery store?
- How much does the median household income vary from county to county in the USA?

GIS technicians are often called on to answer these sorts of questions. Answers to these types of require an understanding of basic statistics. This module describes the fundamental concepts of this statistics, in a light-hearted manner that attempts to express the key ideas of this critical intellectual building block without being overly technical or boring.

Example Application

You are a GIS technician for an environmental consulting firm, Worldwide Environmental Development Organization (WeDo), and your firm has been assigned the task of evaluating the effectiveness of a wetlands restoration project. Wetlands are vital ecological regimes. They serve as breeding grounds for many fish and bird species, and help filter pollutants out of water. Wetlands also tend to be highly sought after as potential locations for development, typically being on the coasts. A current debate in ecology centers around whether or not valuable wetlands in highly urbanized areas can be developed, in return for 'wetland restoration' projects in less urbanized areas. Your firm (WeDo) has been hired by the EPA to evaluate the effectiveness of one such restoration project. Two one-hundred acre test sites (call them A and B), have been set up. Site A is in a wetland, and Site B is in a nearby 'restored' wetland. Each site has been divided into 1 acre pieces, and a catalog of the existence of key species of plant and

animal species has been made. Your mission, if you should accept it, is to determine how similar the two sites are.

Learning Outcomes

The following list describes the expected skills which students should master for each level of training, i.e. Awareness/Competency/Mastery.

Awareness:

Students should be familiar with the basic statistical measures of location, as well as the idea of a sample, histogram and distribution. The meaning and method of calculating the mean, median and mode are covered in this lesson.

Competency:

Students will learn how to calculate the variance and standard deviation of a sample. These idea is expressed in sigma notation, and linked to the ideas of accuracy and precision.

Mastery:

Students will apply the ideas of the mean, median, mode, variance and standard deviation to a set of water samples taking from a polluted river.

Preparatory Units

Recommended:

- **Unit 2 - Demographic Data**
- **Unit 4 - Land Record Data**
- **Unit 5 - Natural Resource Data**
- **Unit 37 - Deriving Regions**

Complementary:

- **Unit 38 - Data Expansion**
- **Unit 40 - Using Reclassification Operators**
- **Unit 41 - Using Boolean Algebra Operators**
- **Unit 51 - Prepraring Digital Presentations**
- **Unit 53 - Communicating About GIS products**

Awareness

Think about the word statistics for a moment. The root of the word is state, and derives from the French word statistique. This is not surprising, since the discipline of statistics arose along with the nation states of the 19th century. This was an exciting time in Europe, if you happened to be a king. In many countries, such as France and England, power was coalescing into single unified goverments. Large state-run industries, such as glass works and clothing mills, funded the armies which were making the kings of European nations into ever-more dominant political forces. People began to be concerned with the state of the state, and began building to develop the intelectual tools to tackle this problem. Policy makers needed information concerning the amount of resources in a country, things that we take for granted knowing today, such as the number of people, or the amount of wheat. Clearly, too much or too little of these resources could be a bad thing for a country in the 18th century. But how can one really tell how much of something exists when we only have incomplete knowledge? This is where statistics comes in. Statistics is a set of mathematical tools we can use to reason effectively in situations where we have incomplete knowledge. Most of the time we have incomplete knowledge. Therefore, understanding a little statistics is vital to understanding the world around us. That is the goal of this chapter.

Statistical Measures of Location

Statistics makes heavy use of spatial metaphors. We have already hinted at this by mentioning the *state of a state*. Another idea used in statistics is that of location. We need to understand, however, that the location's we are discussing are along a number line - *value space*, not real space. As an example, consider the following hypothetical situation:

King Louis the Sixteenth, and his girlfriend, Marie Antionette, are hanging out at Versaille, the royal palace-away-from-home, drinking champagne and playing badmitton. Marie had lept into the air and was just about to enthusiastically return the

shuttlecock when a courtier rushed up to the playing field, all haste and concern, and interrupted with:

Courtier: "Ahem, yer kingship, please excuse this interuption, but I have vital news from the provence of Leon."

Marie: "Off with his head."

Louis: (Who was saved from immanent defeat by the courtier's interuption) "Now Dear, let him speak. Then we'll cut his head off."

Marie: "Off with your head!"

Louis: "Now dear, we've been over this before. I'm Louis the XVI, the semidevine sun king of France, and we can't go around smiting off my head. The craze might catch on, and we'd end up with some dingy form of republic or even worse one of those ..."

Courtier: "Sire. The royal granaries at Leon have been engulfed in flame. The people are starving - they have no more bread!"

Marie: "Let them eat cake!"

Courtier: "But they have no cake!"

Louis: (Picking up a new glass of Champagne) Hmmm, yes, no cake. I wonder. Guards, behead the courtier. Then despatch yourself with all haste to the nearest hamlet, and there enquire within nine houses as to whether or not the peasants there ensconced possess of themselves some certain number of cakes. By this means we shall ascertain the number of cakes possessed by the peasantry.

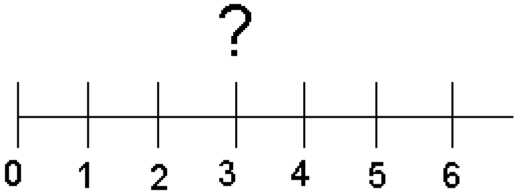
The guards behead the courtier, go to the village, and return with following counts:

House	Cakes
1	3
2	4
3	0
4	0
5	1
6	2
7	3
8	3
9	2

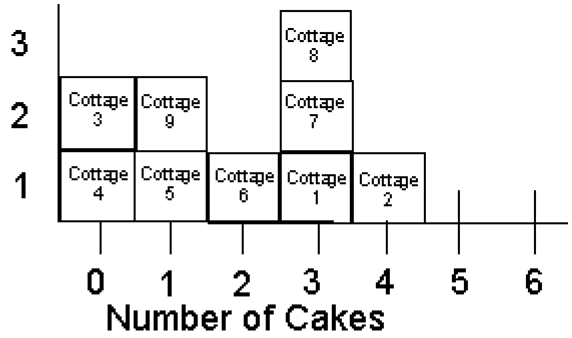
Taken together, the number of cakes in the table above represents a sample of the number of cakes possessed by peasants.

The sample provides us a basis for estimating the number of peasant possessed pastries, based on incomplete information.

Consider the number line below:



Where on this line is the 'location' of peasant possessed pastries? Like any good question, there are actually a few different ways to answer this quandry. One tool that statisticians frequently use to help them understand variables is the HISTOGRAM plot. The histogram has a number line along the x axis and the number of values found along the Y axis. Here's an example, based on the counts reported by the king's guards.



The histogram above depicts what statisticians call a 'distribution'. A distribution associates a given probability with each possible value of a variable. There are three types of location statistics associated with any distribution: the MODE, the MEAN, and the MEDIAN.

The MODE

If we decide that the location is best represented by the number of cakes that appears most often in a distribution, then we could say, 'the peasants have about 3 cakes'. The type of location measurement is called the MODE. The mode of a distribution is similar to the key of a piece of music. The key of a piece of music is generally the note that appears most often, and the mode of a distribution is the value that shows up most frequently. So, for example the mode of the distribution pictured above is '3'. Here is how you can find the mode of a given set of numbers:

- **Find the number of occurences of each possible value**
- **The mode is the value that has the highest number of occurences**

The MEDIAN

The median of a set of numbers is the 'middlemost' value. We can find the middle-most

value by arranging the numbers so that they go from low to high (ascending order), and then looking halfway down the list to find the median value. For example, our sorted list of peasant pastry products would look like:

House	Cakes	
3	0	
4	0	
5	1	
6	2	
9	2	- This is the median value
1	3	
7	3	
8	3	
2	4	

Counting half way down this list, to the fifth number, gives us a house with 2 cakes, therefore we would say that the median number of cakes is 2.

! **Right. You got it. If the number of points is even, then there won't be a middlemost most value. In this case, it is traditional to add the two middlemost values together, and divide them by two. This new number is then used as the median.**

The median may always be calculated in the following manner:

- **Sort the values from lowest to highest (in ascending order)**
- **If the number of values is odd, then the middlemost value is the median**
- **If the number of values is even, then the median value is the sum of the two middlemost values divided by two.**

The MEAN

In statistics the term mean does not denote 'an ill-tempered person who would rather staple you ears to the side of your head than buy you an icecream cone'. Not at all. The statistical MEAN of a set of numbers is just what we would typically refer to as a good old fashioned average. It really is just that simple. The average of a set of numbers is calculated by adding up all the numbers, and then dividing them by the number of numbers. The mean of the distribution of cakes would be calculated as:

```
mean = Sum of the Values
       -----------------
       The Number of Values
```

or, for our set of numbers:

```
mean = 3 + 4 + 0 + 0 + 1 + 2 + 3 + 3 + 2  = 2
       ---------------------------------
                       9
```

We can solve this equation in two steps: by first finding the sum of the values and then dividing this sum by 9, the number of houses visited. The sum of the values is 3 + 4 + 0 + 0 + 1 + 2 + 3 + 3 + 2 = 18. 18 divided by 9 is 2. Therefore the mean number of cakes

found was 2. Voila! Isn't that simple?

But wait. Let's muddy the water just a teensy weensy bit, by re-expressing our formula more mathematically. We could rewrite this equation as:

```
Avg =    Sum( X )
         --------
            N

Where Sum(X) is all the X's added up, and N is the number of X's
```

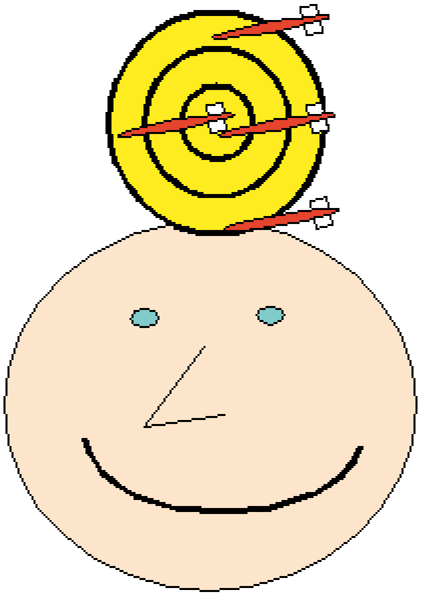
! Right. The average of a set of numbers is the total of all the numbers added together, and divided by number of numbers.

Competency

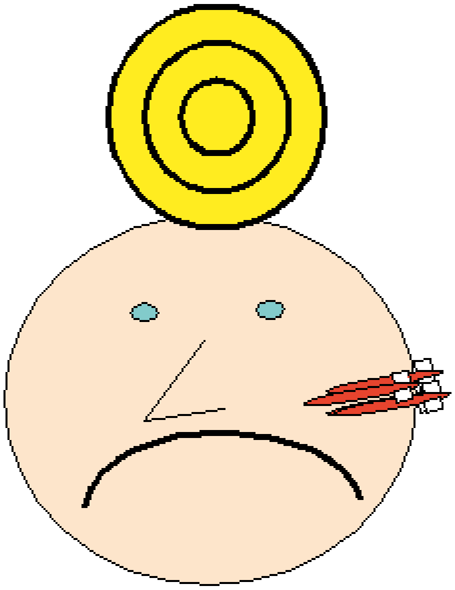
Measures of Dispersion

Most of you will be familar with the difference between accuracy and precision. Accuracy is how well some set of numbers center around a certain target, and precision is how close a set of numbers are to each other. We will use these simple concepts as a bridge to understanding our next major topic: statistical measures of dispersion.

High Accuracy - Low Precision



Low Accuracy - High Precision



Have you ever had a friend who is perfectly consistent in their capacity to really mess something up? I mean, they could be driving through the desert, and they would run into a tree every time. These individuals are precise in their behavior. We all also know

people who are dispersed - running hither or thither with no clear central tendency. In statisticese, w say that a set of numbers (a distribution) is highly dispersed, if the numbers vary greatly from their mean.

Pay attention to that word vary, we'll be revisiting it shortly. Sets of numbers that are tightly clustered around the mean are not dispersed. The table below has some examples:

Dispersed	Not Dispersed
Temperature in Siberia (Hot Summer - Cold Winter)	Temperature in Borneo (Along the Equator - Always the Same)
Big Gambler's Finances (Up and Down)	Student's Finances (Always Low)
Price of Gold	Price of Milk
Popularity of DISCO music	Popularity of Classical Music

And so on. Clearly, things which are dispersed vary from greatly one instance to the next. Things with low variance are always the same. Whoa ... we just snuck a new technical term up on you: VARIANCE. Variance can be defined as how much a distribution varies about the mean. Any guesses as to how we could calculate it? We could just try to add up all the amounts that each number differs from the mean. The problem with this approach is that sometimes these numbers will be postive, and sometimes they'll be negative - and these will always cancel, giving sums equal to zero even for highly dispersed distributions. Statisticians solved this problem by squaring all the differences between each value and the mean. This variance calculation gives you a number in the original units squared. The variance is always positive or zero. A variance of zero means that all the numbers in a sample have exactly the same value - which is also the mean. The exact method for calculating VARIANCE is:

- **Calculate the mean of your set of numbers**
- **Add the square of the differences between the each number and the mean calculated above**
- **Sum these little doggies**
- **Divide this sum by the number of numbers added together in the previous step, minus one**

As descriptions like this become more complicated, mathematical mumbo jumbo becomes more complex. The verbal description from above would look something like:

```
Var =    Sum( SQR( X - Avg ) )
         --------------------
                  N-1

Where the Sum() operator adds everything up,
the SQR() operator squares the differences between each X and
the average of the X's, and N is the number of X's
```

In english, we could express this equation as *the variance equals the sum of the squared deviations from the mean, divided by the number of points minus one. So the variance is a sort of average squared deviation.* Astute readers may be wondering why the variance is divided by n-1, instead of n - the number of points. This is what's known as a 'bias correction'. A bias is a systematic under or over-prediction of a value. Basically, since large values occur very seldomly, these values tend to be missed when calculating the variance based on just a sample. So dividing by N-1, as opposed to N, takes this into account and makes the estimate of the variance unbiased.

Let's calculate the variance of our cake counts. Remember that the mean number of the cakes was 2. We can begin by calculating the squared deviations from the mean.

House	Cakes	Deviations	Deviations-squared
3	0	-2	4
4	0	-2	4
5	1	-1	1
6	2	0	0
9	2	0	0
1	3	1	1
7	3	1	1
8	3	1	1
2	4	2	4

Now we can total the squared deviations:

Total Squared Deviations = 4 + 4 + 1 + 0 + 0 + 1 + 1 + 1 + 4 = 16

Contrast this value with the sum of deviations:

Total Deviation = -2 - 2 - 1 + 0 + 0 + 1 + 1 + 1 + 2 = 0

The total squared deviations is sometimes refered to as the 'variation' of a data set. To get the average 'variance' we divide the variation of a sample by n-1.

VARIANCE = 16 / (n-1) = 16 / (9-1) = 16 / 8 = 2 cakes^2

So the variance of our cake sample is 2 cakes squared. What! Two cakes squared? Yes - that's how the units work out in the above equations. Thinking in terms of cakes squared makes a lot of people (including me) nervous. For this reason, statisticians very frequently use a measure of dispersion called the STANDARD DEVIATION. The standard deviation is just the square root of the variance. Or in mathematical parlance:

```
STANDARD DEVIATION = sqrt(VARIATION)
```

The standard deviation, like the variance, will be zero for samples that all have the same value. The standard deviation will be in the original units of the variable of interest. This is convenient, since we can use it to reason about the expected values that the variable might have. Most values will fall in the range defined by the mean plus or minus the standard deviation. Thus, for our pastry product example, Louis concluded that the peasants, on average, have two cakes, but that some peasants might have none at all.

Mastery

Together, the variance and mean of a set of numbers tell us a great deal: both the location and level of dispersion of a data set. For example, examine this linked image: china.jpg. This image shows a set of weather stations in Asia. The weather stations are stars, and the color of each station represents the temperature of china. Can you see how it gets colder as you go farther north? Below the map you'll find a histogram, expressed in degrees celcius, so zero is freezing. The mean of this data set is below freezing (-2.5 degrees C) - so we'd want to bring our parkas - and the standard deviation is 10 degrees, so it is a lot colder in the north than in the south.

Now, imagine that you are Dr. Doolittle, a renowned scientist, called in to assess the dangers posed by a lead pipe factory, Romullus and Remus Incorporated. R&R inc is a new factory, located in nothern California along a small tributary of the Trinity river. An ecological group has has recently reported increased levels of lead in the tributary, which would be a bad thing, since lead is a poison. The Environmental Protection Agency sends you out as leader of a crack team of researchers to study this matter. On site, you discover that the ecologist group has been taking water samples for the last seven years. Their data looks like this:

Year	Lead (in parts per billion)
1995	222
1996	120
1997	222
1998	176
1999	473
2000	538
2001	456

Calculate the mean, median, and mode for these data points. Which measure is most appropriate for assessing public safety requirements? Calculate the variance and standard deviance of the amount lead of lead. The R&R INC plant plant was built in 1999. Does it seem reasonable to attribute the increase in lead amounts to random chance?

When you have your answers, click here to see if you're correct.

Follow-up Units

Resources

[Outdated links have been removed.]

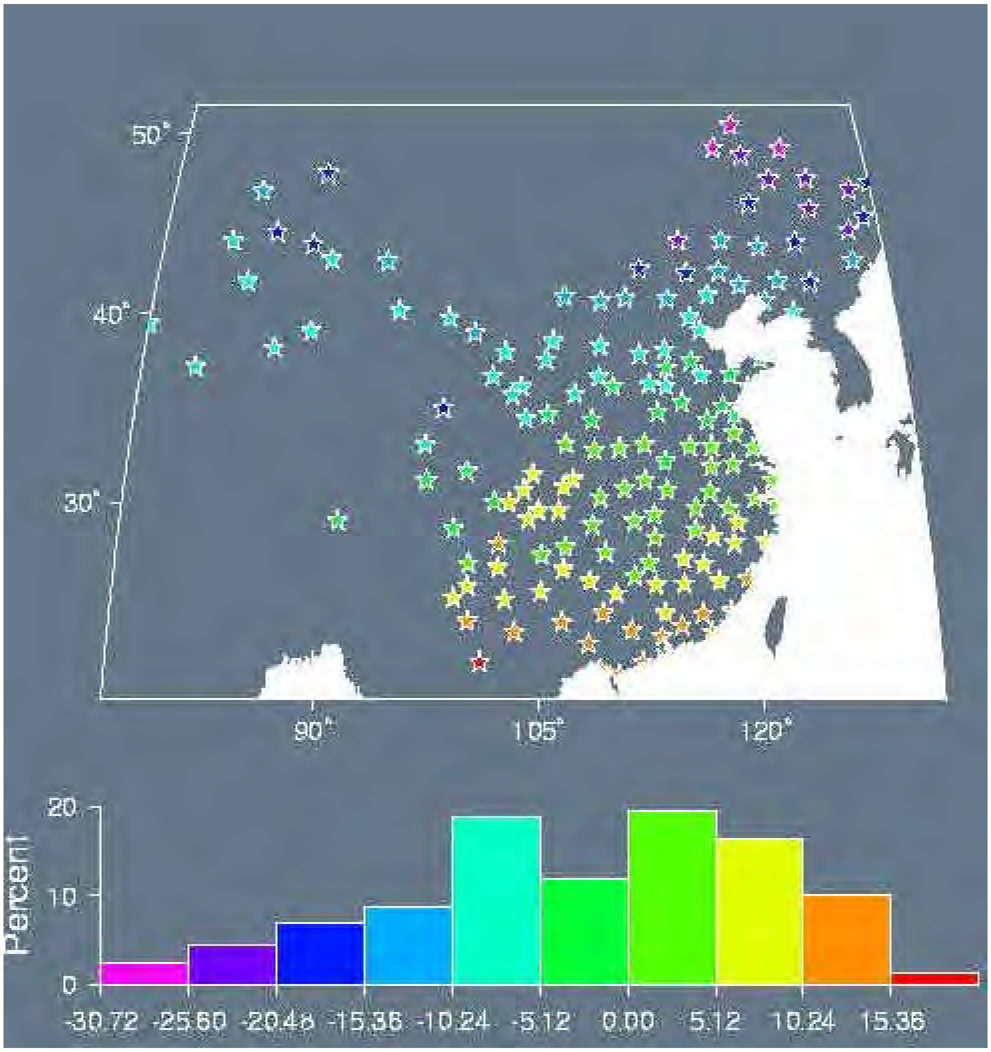
Surfstat This is a great website full of java applets and great descriptions. Highly recommended.

Solveit This page more example of mean, median and mode examples.

Sampling This page has some good discussion and example of statistical samples.

SFU This page from Simon Frazer University provides a good summary of the topics we've discussed.

Created: May 14, 1997. Last updated: October 5, 1998.


50°
40°
30°
90°
105°
120°
Percent
20
10
0
-30.72
-25.60
-20.48
-15.36
-10.24
-5.12
0.00
5.12
10.24
15.36

china.jpg

Answers for Unit 39

1. To calculate the mean for this data set we simply:

    - Sum up the numbers

      Sum = 222 + 120 + 222 + 176 + 473 + 538 + 456 = 2207 ppb lead

    - And divide by the number of numbers

      Avg = 2315 / 7 = 315

---

2. To calculate the median of these numbers, we can:

    - Sort the numbers from smallest to largest

      120
      176
      222
      222
      456
      473
      538

    - And then select the middle number:

      120
      176
      222
      222 <--- the median
      456
      473
      538

      So the median for this data set is 222 ppb lead

---

3. To calculate the mode of this data, we can:

    - Count the number of occurences for each number

#	Occurences
120	1
176	1
222	2

456	1
473	1
538	1

- And then select the number that occurs most frequently:

#	Occurences
120	1
176	1
222	2 <------ the mode
456	1
473	1
538	1

So the mode for this data set is the same as the median: 222 ppb lead

---

Which measure seems the most appropriate for evaluating a public safety risk? Each measure has its advantages. The mode tells which value is most frequent, which is certainly important for accessing risk, since this is the value most commonly encountered. The median is useful to, since it gives an estimate for the 'middlemost' value that is not overly-influenced by extreme values. The median and the mode for this data set are the same (222 parts per billion of lead) and are quite lower than the mean (315 ppb lead). We might be tempted to believe that some extreme high values contributed to the difference between the median and mean values. Look again at the time-series of values:

Year	Lead (in parts per billion)
1995	222
1996	120
1997	222
1998	176
1999	473
2000	538
2001	456

Notice how all the higher values come in the later years? This does seem to suggest (not prove!) that there has been a shift in the levels of lead since 1999. This apparent shift could be just that - mere appearance - and the diffence between the pre and post plant values could simply be a random occurence. One way of thinking quantitatively about these differences is to use the variance and the standard deviation to determine how large a typical difference might be. We'll do this in the next section:

---

4. To calculate the variance of this data, we can:

    - Calculate the difference between each value and the mean

      222 - 315 = -93
      120 - 315 = -195

222 - 315 = -93
176 - 315 = -139
473 - 315 = 158
538 - 315 = 223
456 - 315 = 141

- And then square each of these values:

  -93 * -93 = 8649
  -195 * -195 = 38025
  -93 * -93 = 8649
  -139 * -139 = 19321
  158 * 158 = 24964
  223 * 223 = 49729
  141 * 141 = 19881

- And then sum these squared values:

  Sum = 8649 + 38025 + 8649 + 19321 + 24964 + 49729 + 19881 = 169218 ppb lead

- And then divide this sum by the number of numbers minus one:

  Variance = 169218 / (7-1) = 28203 ppb lead

The variance can be difficult to interpret, especially since is units will be the square of the original variable's units. A more easily understood statistic is the standard deviation which we can calculate by:

> Taking the square root of the variance:
>
> Standard Deviation = sqrt(28203) = 168 ppb lead

Hmmmmmmm. The mean lead level for the first four years was (222+120+222+176)/4 = 185, and the mean lead level for the last 3 years (after the plant was in operation) was (473+538+456)/3 = 489. So the amount of lead in the years following the building of the plant was more than 2 and a half times it's pre-1999 levels. Furthermore, the size of the jump between pre and post plant means was 489 - 185 = 304 parts per billion of lead, or almost 2 whole standard deviations. This is unlikely to occur totally by chance, and we suggest that further reasearch into the Romullus and Remus lead pipe factory is in order.