MODULE 1

WHY STATISTICS IS IMPORTANT

Objectives:

1. Describe the two major reasons why statistics is important for understanding populations.
2. Define natural and sampling variability.
3. Describe "difficulties" in making conclusions about population caused by sampling variability.
4. Define "statistics" (as a field of study).
5. Appreciate the importance of statistics in scientific inquiry.

Contents

1.1 Realities

The city of Ashland performed an investigation in the area of Kreher Park (Figure 1.1) when considering the possible expansion of an existing wastewater treatment facility in 1989. The discovery of contamination from what was believed to be creosote waste in the subsoils and ground water at Kreher Park prompted the city to abandon the project. A subsequent assessment by the Wisconsin Department of Natural Resources (WDNR) indicated elevated levels of hazardous substances in soil borings and ground water samples and in the sediments of Chequamegon Bay directly offshore of Kreher Park. In 1995 and 1999, the Northern States Power Company conducted investigations that further defined the area of contamination and confirmed the presence of specific contaminants associated with coal tar wastes. This site is now listed as a superfund site and is being given considerably more attention.[1]

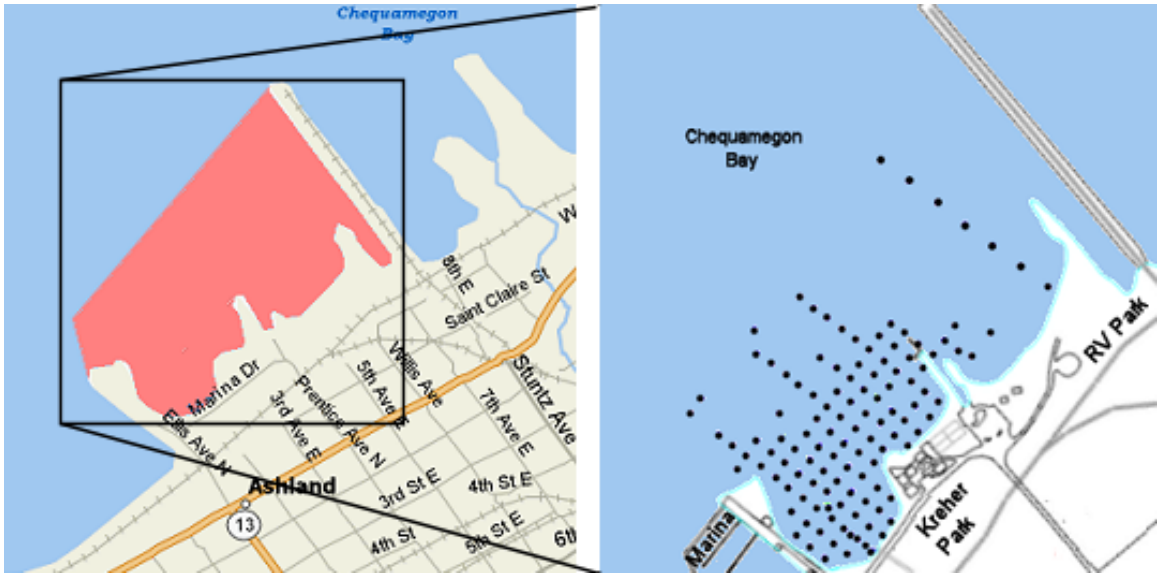


Figure 1.1. Location of the Ashland superfund site (left) with the location of 119 historical sediment sampling sites (right).

The WDNR wants to study elements in the sediment (among other things) in the entire 3000 m^2 area shaded in Figure 1.1. Is it physically possible to examine every square meter of that area? Is it prudent, ecologically and economically, to examine every square meter of this area? The answer, of course, is "no." How then will the WDNR be able to make conclusions about this entire area if they cannot reasonably examine the whole area? The most reasonable solution is to sample a subset of the area and use the results from this sample to make inferences about the entire area.

Methods for properly selecting a sample that fairly represents a larger collection of individuals are an important area of study in statistics. For example, the WDNR would not want to sample areas that are only conveniently near shore because this will likely not be an accurate representation of the entire area. In this example, it appears that the WDNR used a grid to assure a relatively even dispersal of samples throughout the study area (Figure 1.1). Methods for choosing the number of individuals to select and how to select those individuals are discussed in Module 3.

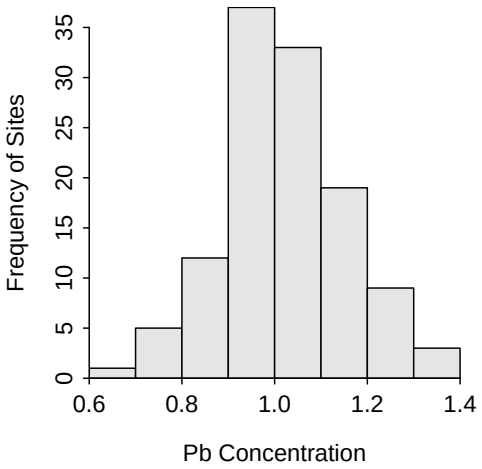
Suppose that the WDNR measured the concentration of lead at each of the 119 locations shown in Figure 1.1. Further suppose that they presented their results at a public meeting by simply showing the list of

[1] More information at the EPA and the WDNR websites.

lead concentration measurements (Table 1.1).[2] Is it easy to make conclusions about what these data mean from this type of presentation? Instead, suppose that the scientists came to the meeting with a simple plot of the frequency of observed lead concentrations and brief numerical summaries (Figure 1.2). With this presentation one can easily see that the measurements were fairly symmetric with no obviously "weird" measurements and ranged from as low as 0.67 $\mu g \cdot m^{-3}$ to as high as 1.36 $\mu g \cdot m^{-3}$ with the measurements centered on approximately 1.0 $\mu g \cdot m^{-3}$. These summaries will be discussed in detail in Module 5. However, at this point, note that statistical methods are important for distilling or summarizing large quantities of data into graphs or numerical summaries from which it is easier to identify meaning from the data.

Table 1.1. Lead concentration ($\mu g \cdot m^{-3}$) from 119 sites in Kreher Park superfund site.

0.91	1.09	1.00	1.09	1.06	0.98	0.98	0.94	0.89	1.09	0.91	1.06	0.81	0.90	1.21
1.03	0.95	1.14	0.99	0.99	0.96	1.13	0.84	1.03	0.86	0.98	1.04	0.91	1.27	0.90
0.87	1.23	1.12	0.98	0.79	1.10	1.06	1.09	0.73	0.81	1.18	0.92	0.82	1.11	0.97
1.24	1.06	1.09	0.78	0.94	1.08	0.91	0.98	1.22	1.04	0.77	1.18	0.93	1.14	0.94
1.05	0.91	1.14	0.93	0.94	0.90	1.05	1.36	1.02	0.93	1.09	1.17	0.91	1.06	0.95
0.88	0.67	1.12	1.06	0.99	0.89	0.83	0.99	1.33	1.00	1.05	1.11	1.01	1.25	0.96
1.07	1.17	1.01	1.20	1.17	1.05	1.21	1.10	1.07	1.01	1.16	1.24	0.86	0.90	1.07
1.11	0.99	0.70	0.98	1.11	1.12	1.30	1.00	0.89	0.91	0.95	1.08	1.02	0.93	

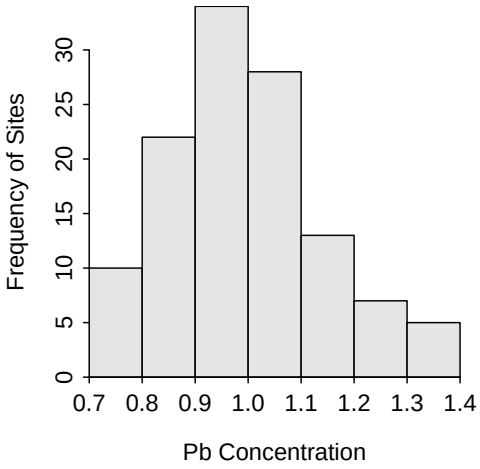


	Pb
Min.	0.67
1st Qu.	0.925
Median	1.01
Mean	1.016
3rd Qu.	1.1
Max.	1.36

Figure 1.2. Histogram and summary statistics of lead concentration measurements ($\mu g \cdot m^{-3}$) at each of 119 sites in Kreher Park superfund site.

A critical question at this point is whether or not the results from the one sample of 119 sites perfectly represents what the results would be for the entire area. One way to consider this question is to examine the results obtained from another sample of 119 sites. The results from this second sample (Figure 1.3) are clearly, though not radically, different from the results of the first sample. Thus, it is seen that any one sample from a large area will not perfectly represent the area. Furthermore, it is observed that two different samples give two different results which will likely lead to two different, albeit generally only slightly different, conclusions.

[2] These are hypothetical data for this site.



	Pb
Min.	0.71
1st Qu.	0.89
Median	0.98
Mean	0.9919
3rd Qu.	1.08
Max.	1.39

Figure 1.3. Histogram and summary statistics of lead concentration measurements ($\mu g \cdot m^{-3}$) at each of 119 sites (different from the sites shown in Figure 1.2) in Kreher Park superfund site.

The results of two different samples do not perfectly agree because each sample contains different individuals, and no two individuals (sites in this example) are exactly alike. The fact that no two individuals are exactly alike is **natural variability**, because of the "natural" differences that occur among individuals. The fact that the results from different samples are different is called **sampling variability**. If there was no natural variability, then there would be no sampling variability. If there was no sampling variability, then the field of statistics would not be needed because a sample (even of one individual) would perfectly represent the larger group of individuals. Thus, understanding variability is at the core of statistical practice. Natural and sampling variability will be revisited continuously throughout this course.

Δ **Natural Variability**: The fact that no two individuals are exactly alike.

Δ **Sampling Variability**: The fact that the results (i.e., statistics) from different samples are different.

This may be a bit unsettling! First, it was shown that an entire area or all of the individuals of interest cannot be examined. It was then shown that a sample of individuals from the larger set did not perfectly represent all of the individuals. Furthermore, each sample is unique and will likely lead to a different conclusion. These are all real and difficult issues faced by the practicing scientist and considered by the informed consumer. However, the field of statistics is designed to "deal with" these issues such that the results from a relatively small subset of measurements can be used to make conclusions about the entire collection of measurements.

$\diamond$ **Statistics provides methods for overcoming the difficulties caused by the requirement of sampling and the presence of sampling variability.**

1.2 Major Goals or Purposes of Statistics

The field of statistics has two primary purposes, which were illustrated in the Kreher Park example above. First, statistics provides methods to summarize large quantities of data into concise and informative numerical or graphical summaries. For example, it was easier to discern the general underlying structure of the lead measurements from the statistics and histograms presented in Figures 1.2 and 1.3 than it was from the full list of lead measurements in Table 1.1. Second, statistical methods allow inferences to be made about all individuals (i.e., a population) from a few individuals (i.e., a sample). Population and sample are defined more completely in Section 2.1.

◇ **Statistics, as a field of study, is used to (1) summarize large quantities of data and (2) make inferences about populations from samples.**

1.3 Definition of Statistics

Statistics is the science of collecting, organizing, and interpreting numerical information or data (Moore and McCabe 1998). People study statistics for a variety of reasons, including (Bluman 2002):

1. They must be able to read and understand the statistical studies performed in their field. To have this understanding they must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in those studies.
2. They may need to conduct research in their field. To accomplish this they must be able to design experiments and samples; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. They must also be able to communicate the results of the study.
3. They also need to be better consumers of statistical information.

The science of statistics permeates a wide variety of disciplines. Moore and McCabe (1998) state:

> The study and collection of data are important in the work of many professions, so that training in the science of statistics is valuable preparation for a variety of careers. Each month, for example, government statistical offices release the latest numerical information on unemployment and inflation. Economists and financial advisers, as well as policy makers in government and business study these data in order to make informed decisions. Doctors must understand the origin and trustworthiness of the data that appear in medical journals if they are to offer their patients the most effective treatments. Politicians rely on data from polls of public opinion. Business decisions are based on market research data that reveal customer tastes. Farmers study data from field trials of new crop varieties. Engineers gather data on the quality and reliability of manufactured products. Most areas of academic study make use of numbers, and therefore also make use of the methods of statistics.

Δ **Statistics**: The science of collecting, organizing, and interpreting numerical information or data.

Review Exercises

1.1 There are 1499 lakes in Ashland, Bayfield, and Douglas counties of Wisconsin. However, only 605 of these are named. A random sample of named lakes from this population is extracted with the following R code:

```
> library(NCStats)
> named <- filterD(ABDLakes,named)
> srsdf(named,n=50,vars=c("county","area"))
```

Use this code and some hand (or calculator) calculations to answer the questions below. *Answer*

(a) Extract a sample of n=50 lakes with the code above. Compare the sizes (area in acres) of the first three lakes. This is an example of what type of variability?

(b) Compute the proportion of lakes in your sample that are from Bayfield County.

(c) Extract another sample of n=50 lakes and compute the proportion of lakes that are from Bayfield County? Compare your two proportions. This is an example of what type of variability?

(d) Of the named lakes in the three counties, 346 are from Bayfield County. Was the proportion of lakes from Bayfield County in either of your samples equal to the proportion of all named lakes that were from Bayfield County? Were you surprised? Why or why not?