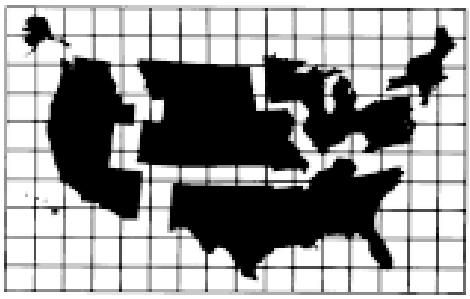
HAZARDOUS SUBSTANCE RESEARCH CENTERS

ENVIRONMENTAL SCIENCE AND TECHNOLOGY BRIEFS FOR CITIZENS



by K.J. McGaughey

Understanding Statistics

Interpreting environmental conditions through math

This publication is published by the Hazardous Substance Research Centers as part of their Technical Outreach Services for Communities (TOSC) program series of Environmental Science and Technology Briefs for Citizens. If you would like more information about the TOSC program, contact your regional coordinator:

Northeast HSRC
New Jersy Institute of Technology
Otto H. York CEES
138 Warren St.
Newark, NJ 07102
(201) 596-5846

Great Plains/Rocky Mountain HSRC
Kansas State University
101 Ward Hall
Manhattan, KS 66506
(800) 798-7796

Great Lakes/Mid-Atlantic HSRC
A-124 Research Complex-Engineering
Michigan State University
East Lansing, MI 48824
(800) 490-3890

South/Southwest HSRC
Environmental Science & Technology Program
Georgia Tech Research Institute
229 Baker Building
Atlanta, GA 30332
(404) 894-7428

Western Region HSRC
Oregon State University
210 Strand Agriculture Hall
Corvallis, OR 97331-2302
(800) 653-6110

K-STATE
Kansas State University

Acknowledgment: Although this article has been funded in part by the U.S. Environmental Protection Agency under assistance agreement R-819653, through the Great Plains/Rocky Mountain Hazardous Substance Research Center, it has not been subjected to the agency's peer and administrative review and, therefore, may not reflect the views of the agency. No official endorsement should be inferred.

The condition of air, water, and soil results from many natural and human-made factors, and varies greatly from point to point across the earth. Statistical methods use measurements and numbers to help us understand these environmental conditions and to predict their effects on plants and animals. If a problem is found to exist, statistical methods can help us decide what corrective action might work best and, afterwards, to confirm the problem has been solved.

What is statistics?

Statistics is the study of how to collect, organize, analyze, and interpret numerical information. It provides the framework for decision making when testing hypotheses formulated through the scientific method. The goal of statistics is to learn something about a population based on measurements taken from a sample. A **population** is a group of individuals, objects, or units about which we wish to know something. A **sample** is a collection of members of the population on which measurements are to be taken. The word "sample" is often used to refer to the measurements themselves. Some examples of populations and samples are as follows:

- The department of tropical agriculture is doing a study of the weights of pineapples in an experimental field. A random collection of 100 pineapples is taken from the field. The population is the weights of all pineapples in the field. The 100 weights form a random sample from the population of all weights.

- Federal research scientists wish to determine if a certain tract of land is free from toxic chemicals. The tract is divided into 100, ten-foot by ten-foot squares; 25 of these are randomly selected for testing. The population is the total land tract the scientists are interested in. The sample is the 25, ten-foot by ten-foot squares selected for testing.

- A research scientist wishes to determine if a certain proportion of prairie dogs in the state of Colorado carry bubonic plague. One hundred prairie dogs are caught and tested for the disease. The population is all of the prairie dogs in Colorado. The sample is the 100 prairie dogs which were caught and tested.

What do statisticians do?

Statisticians draw inferences or conclusions about a population based on what they observe in the sample. Sometimes we do not have access to the entire population and at other times the difficulties of working with an entire population are prohibitive. The benefit of statisticial methods is that it allows us to draw conclusions about populations based on only the information from samples. When using statistical methods, there will always be a margin of error due to using a sample, rather than the entire population. The relationship between sample and population is crucial; if a sample is representative of only a portion of the population, then it is a biased sample. Inferences drawn from a biased sample

Definition of terms

Population: a group of individuals, objects, or units about which we wish to know something.

Sample: a collection of members of the population on which measurements are to be taken. The word "sample" is often used to refer to the measurements themselves.

Bias: not actually measuring what one hoped to measure.

Random Sample: a sample which is representative of the entire population; i.e., each member of the population has an equal chance of being included in the sample.

Parameter: a number or characteristic of the population.

Statistic: a number or characteristic of the sample.

may be dangerously misleading. To avoid biases in a sampling procedure, samples should be drawn at random from the population. Most mathematical results in statistics pertain to random samples.

Continuing with the prairie dog example: One way to sample would be to find one prairie dog town in the northeastern corner of the state and capture 100 prairie dogs for our bubonic plague testing. A second way to sample would be to capture one prairie dog out of each of 100 randomly chosen prairie dog towns from across the entire state. The first sampling method gives us biased results if we try to draw conclusions about all prairie dogs in Colorado. We sampled from one town, thus any conclusions we draw are applicable to that prairie dog town only. Whereas in the second sampling scenario, because we sampled randomly from across the whole state, any conclusions we draw may be applied to the entire population of prairie dogs in the state of Colorado.

What is hypothesis testing?

A statistical hypothesis, or a hypothesis, is a statement about some characteristic of a population. Such a characteristic is called a population parameter. In a statistical hypothesis test, the hypothesis will be accepted or rejected on the basis of information extracted from data. Two hypotheses are used in a statistical hypothesis test. The first is called the null hypothesis. This is the assertion held as true until we have enough statistical evidence to conclude otherwise. It is usually denoted as H_O. The second is called the alternative hypothesis. This is the assertion of all situations not covered by the null hypothesis. It is usually denoted as H_A.

For our prairie dog example, possible null and alternative hypotheses are as follows:

H_O: The proportion of prairie dogs in Colorado carrying bubonic plague is less than or equal to (<or=)0.10, or 10%.

H_A: The proportion of prairie dogs in Colorado carrying bubonic plague is greater than (>) 0.10.

When learning about hypothesis testing, it is helpful to consider the example of a jury trial. By convention, the accused is considered innocent until proven guilty in this country, so the null hypothesis, H_O, is defined to be "not guilty," and the alternative hypothesis, H_A, is defined to be "guilty." If the accused is innocent, but the jury finds she/he is guilty, they have sent an innocent person to prison and made an error. If the accused is guilty of the crime, but the jury finds she/he is not guilty, or innocent, they have let a guilty person go free and again made an error. A good statistical procedure is one that will minimize the possibility of making mistakes like these in our decisions. Statisticians design procedures that will, while not eliminate error, at least reduce the chances of making an error. We can usually only discuss the probability of making an error because we do not know for sure if an error has been made. (This would require knowing the characteristics of our population in advance, in which case the hypothesis test would be unnecessary.)

Assume for the prairie dog example that if the proportion of prairie dogs infected by the bubonic plague is less than 0.10, there is no danger of an outbreak and no remediation is required. But if this proportion is greater than 0.10, there is a danger of an outbreak and we must spend time and money to control it. When performing a hypothesis test for this scenario, we make an error when we have decided the proportion of prairie dogs in Colorado carrying the plague is greater than or equal to 0.10, when in fact it is less than 0.10. In which case we might take action to reduce the level below 0.10 when, in fact, that cost was not necessary. If we

Sample Size	Number Infected	Proportion Infected	80 % Confidence Interval	90 % Confidence Interval	95 % Confidence Interval
30	4	0.13	0.13 ± 0.079 or (0.0514, 0.2086) width = 0.1572	0.13 ± 0.101 or (0.0290, 0.2310) width = 0.2020	0.13 ± 0.120 or (0.0097, 0.2503) width = 0.2406
60	8	0.13	0.13 ± 0.056 or (0.0744, 0.1856) width = 0.1112	0.13 ± 0.071 or (0.0586, 0.2014) width = 0.1428	0.13 ± 0.085 or (0.0449, 0.2151) width = 0.1702
120	16	0.13	0.13 ± 0.039 or (0.0907, 0.1693) width = 0.0786	0.13 ± 0.051 or (0.0795, 0.1805) width = 0.1010	0.13 ± 0.060 or (0.0698, 0.1902) width = 0.1204

Figure 1. Confidence interval. Increasing the sample size of prairie dogs tested decreases the interval width and margin of error.

decide the proportion of prairie dogs in Colorado carrying the plague is less than 0.10 when, in fact, this proportion is actually greater than or equal to 0.10, we have again made an error. In this case we would feel there is not a problem when, in fact, there really is a danger. To summarize: one type of error costs time and money; the other possibly costs human lives.

What is a confidence interval?

A confidence interval is a range of numbers based on the data of the sample believed to include the population parameter we are interested in. Put simply, it is an interval which we believe captures the true value of the population parameter we are studying. Think of a confidence interval as a net. The larger it is, the more likely it is to capture the unknown population parameter. The smaller it is, the less likely it is to capture the unknown population parameter. The certainty or the confidence we have that this interval contains the unknown population paramenter is also associated with the interval. With a 95% confidence interval, we are 95% certain that the interval contains the population paramter in question. With an 80% confidence interval, we are 80% certain the interval contains the population parameter in question. Thus, a 95% confidence interval must necessarily be longer than an 80% confidence interval to have a greater chance of catching the unknown population parameter. The more certain we wish to be, the larger the interval must be, or the larger our sample size must be. Typically, when results of polls and surveys are published, the confidence interval will be reported as the sample statistic plus or minus a term called the margin of error. The margin of error is the error due to using a sample instead of the entire population.

Figures one and two illustrate how increasing the sample size of praire dogs tested decreases our interval width and margin of error, which in turn increases how certain we are of our estimation of the proportion actually infected. For a sample size of 30, we found four prairie dogs to be infected, and confidence intervals were constructed. Our 95% confidence interval was found to include a range of infection values from .0097 to 0.2503. We interpret this to mean we are 95% certain that this interval contains the population proportion of prairie dogs infected with bubonic plague; i.e., we are 95% confident that the proportion of infected praire dogs in the state of Colorado is 13% +/- 12%. Similarily, an 80% confidence interval for the same sample size and proportion of infected prairie dogs ranges from 0.0514 to 0.2086. We are 80% certain the population of infected prairie dogs is 13% +/- 7.9%. The higher the degree of confidence we wish to have, the larger our interval will be. If we wish to maintain a high degree of confidence, and also have a narrower interval width, or smaller margin of error, we must increase the sample size.

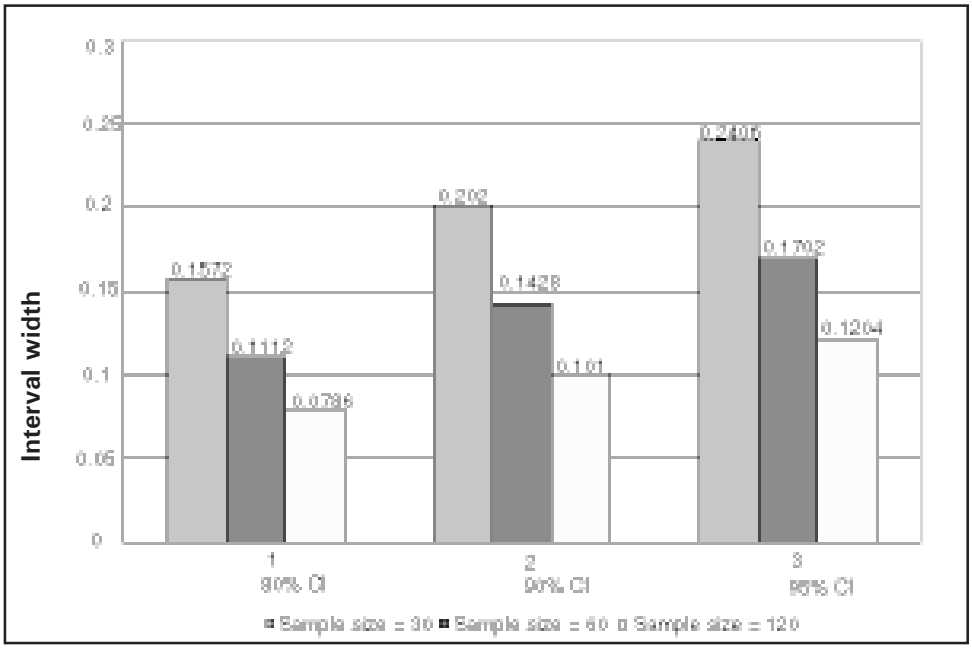


Figure 2. Comparison of confidence interval widths.

How is statistics, or statistical methods, a part of the scientific method?

Statistics plays an active role in the steps of the scientific method. Specifically, statistics helps to formulate a hypothesis, then gives the scientist the tools to test this hypothesis. Depending on the results of the statistical analysis, the scientist will be able to confirm or reject the hypothesis. In the event the hypothesis is rejected, the statistical data may offer a good idea of a new hypothesis for further investigation.

Let's see how statistics and/or statistical methods are incorporated in the scientific method in a hypothetical scenario.

Step 1: (Observation) Several cases of bubonic plague have been reported in southeast Colorado.

Step 2: (Deduction) From the body of scientific knowledge, we know that prairie dogs are known carriers of the plague. For this example we will also assume that we know there is no danger of an outbreak if the proportion of prairie dogs infected is less than 0.10 or 10%.

Step 3: (Hypothesis) There is no potential health risk from prairie dogs.

Step 4: (Testing) We want to draw conclusions about the entire population of prairie dogs in Colorado, but to test each and every prairie dog is not feasible. Thus a method of random sampling is used. Data is collected on samples of prairie dogs and used to test a statistical hy-

pothesis. For example, our null hypothesis may be that the proportion of infected prairie dogs is less than 0.10. It is in this step of the scientific method where statistics plays its greatest role.

Step 5: (Observation) Using confidence intervals and hypothesis testing, the results may indicate the proportion of infected prairie dogs is greater than 0.10. In this case, we would go back to step 3, reformulate our hypothesis, and repeat the testing procedure. If the results do not lead to a decisive conclusion, the researcher reformulates the hypothesis and runs the experiment again.

Conclusion

Statistical methods enable scientists to study numerical information regarding the environment. Statistics also assist decision makers in solving problems related to environmental conditions. Statisticians employ several procedures to minimize the possibility of error in applying statistical methods. The number of samples taken can help increase the reliability of the information revealed by analyzing statistical data. Statistics are an integral part of the scientific method, which assists scientists in formulating and testing hypotheses.

References:

Understandable Statistics; Brase/Brase ©1995 by D. C. Heath and Company.

Dr. George Milliken, Kansas State University, Department of Statistics.

Dr. Tom Loughin, Kansas State University, Department of Statistics.

Dr. James Higgins, Kansas State University, Department of Statistics.

■ ■ ■

ABOUT THE AUTHOR: Karen McGaughey has a B.S. in Chemistry/Education from Kansas State University (KSU) and is currently working on her M.S. in Statistics at KSU.

K-STATE Kansas State University

Great Plains/Rocky Mountain HSRC
Kansas State University
101 Ward Hall
Manhattan, KS 66506
(800) 798-7796