Chapter

1

Statistics:
The Big Ideas and Essential Understandings

The need for the discipline of statistics arises from the "omnipresence of variability in data," in the words of Cobb and Moore (1997, p. 801), who describe the special focus of statistics in the following way:

> The focus on variability naturally gives statistics a particular content that sets it apart from mathematics, itself, and from other mathematical sciences, but there is more than just content that distinguishes statistical thinking from mathematics. Statistics requires a different kind of thinking, because data are not just numbers, they are numbers with a context. In mathematics, context obscures structure. In data analysis, context provides meaning. (p. 801)

This description highlights two important distinctions between statistical and mathematical thinking. First, statistical thinking focuses on engaging in a process that is centered on understanding and describing variability in data. Without variability, there would be no need for statistics, since everything would be the same. Second, context plays a critical role in the practice of statistics. Creating a graph or computing a mean without context is not statistics. In the practice of statistics, the context of the problem under study drives the method of data collection, the analysis of the data, and the interpretation of results.

The purpose of this book is to introduce the big ideas and essential understandings that teachers need to know to help middle school students advance their statistical thinking. These concepts are vital for teachers to understand—especially in light of the release of the Common Core State Standards for Mathematics (Common Core State Standards Initiative 2010), which have increased the amount of statistical content that students in grades 6–12 are

expected to master and the level of sophistication in statistical thinking that they are expected to demonstrate.

Statistics as a Problem-Solving Process

The statistical problem-solving process described in the American Statistical Association's *Guidelines for Assessment and Instruction in Statistics Education* (GAISE Framework) (Franklin et al. 2007) will serve as the foundation for the concepts and illustrations that are presented throughout this book. This investigative process involves four interrelated components:

1. Formulating a statistical question—a question that can be addressed with data
2. Designing a plan for collecting useful data, implementing the plan, and collecting data
3. Analyzing the data—creating and exploring various representations of the distribution to identify and describe patterns in the variability in the data and summarize various features of the distribution
4. Interpreting the results—providing a statistical answer to the question posed that takes the variability in the data into account.

To practice statistics, one must engage in all components of the statistical problem-solving process.

Statistical problem solving begins with a question. For example, statistics is used to address questions such as the following:

Question 1: In a presidential election, do potential voters support a particular candidate?

Question 2: How do the annual salaries for men and women in similar occupations compare?

Addressing either of these questions requires data. Data consist of observations or measurements on a *variable*. In statistics, a variable is a characteristic that may be different from one individual to another. For example, we do not expect all voters to give the same response regarding a particular candidate, and different people in the same occupation are likely to have different annual salaries.

To address the first question, a group of potential voters would be selected, and each person would be asked whether or not he or she would vote for a particular candidate. The variable in this case is *categorical*, since the possible choices for a response are "yes," "no," and "not sure," and on the basis of the response given, each participant is placed into one of these three categories. It is common to summarize data such as these by the number of responses in each category or the percentage of responses in each category.

To address the second question, we would first identify an occupation that employs both men and women. Groups of female and male employees would be selected and the variable "annual salary" recorded for each employee. Annual income is a *quantitative* variable, since the possible values are numbers on which it is appropriate to perform arithmetic operations. For example, it makes sense to determine the mean annual salary—the arithmetic average—as a numerical summary for those individuals surveyed.

Note that some variables result in values that are numbers but nonetheless are not quantitative variables. For example, if each person in a survey is asked her or his zip code, the resulting data appear to be numbers, but the mean zip code does not have a meaningful interpretation. Zip codes are digit combinations serving merely as labels. In this case, zip code is a categorical variable and provides information about the state and city in which an individual resides.

Ideally, the data on the selected units are representative of the data for the larger group of interest (e.g., all potential voters or all people employed in this occupation). In each case, the larger group of interest is called the *population*, and the group selected is called a *sample*. An important component of statistical problem solving is designing and implementing a sampling strategy that tends to produce samples that are representative of the population with regard to the question under study. The most common strategy employed to improve the likelihood of obtaining a representative sample is to incorporate *randomness* into the sample selection procedure. A sample that is representative of the larger population is especially important if we desire to generalize results from a sample to the population.

It is common in statistics to report numerical summaries of data, and it is important to distinguish between numerical summaries based on a population and numerical summaries based on sample data. Numerical summaries of a population are called *parameters*; numerical summaries of a sample are called *statistics*. For example, suppose that 45% of all potential voters support a particular candidate. This percentage is a *parameter*. If a sample of 200 potential voters is selected and 43% of those in the sample support the candidate, then this percentage is a *statistic*. Notice that the percentage in this sample is not the same as the percentage in the population.

In statistics, data are collected to address a question. For example, to compare salaries for men and women in similar occupations, a sample of data on annual salary might be collected. The data in the sample would vary, and identifying patterns in the variation would be useful for providing an answer to the question. As Wild (2006) states, "Statisticians look at variation though a lens which is 'distribution'" (p. 11).

The variation in data within a sample is described by the *sample distribution*. Representations of the sample distribution provide information on what data values occurred and how often each value occurred. The distribution can be summarized by a graph or a table or with numerical summaries. The goal of data analysis is to explore various representations of the distribution to—

- gain a better understanding of the variability in the data;
- identify patterns or trends in the variability in the data; and
- summarize important features of the distribution.

Results from a statistical study are based on data that vary, and their interpretation must allow for this variability. The annual salaries for people with similar occupations vary, and an interpretation of the data on annual salaries must include a description of this variability. To address issues related to differences between the annual salaries of men and women with similar occupations, we would compare their respective sample distributions. Such a comparison involves identifying aspects of the distributions that are similar and aspects that are different.

To summarize what we have said so far, we emphasize that statistics is a problem-solving process that begins with a statistical question. Data are required to address a statistical question. The nature of the analysis and interpretation of the results depend on the question posed, the type of data collected, and the manner in which the data have been collected. It is important to note that variability plays a role in each component of the statistical problem-solving process described in the GAISE Framework.

Employing this statistical problem-solving process is critical in developing an understanding of the big ideas in statistics. Discussion of the big ideas and their associated essential understandings throughout this book will further clarify many of the topics discussed up to this point. The big ideas and all the associated understandings are identified as a group below to give you a quick overview and for your convenience in referring back to them later. Read through them now, but do not think that you must absorb them fully at this point. The chapter will discuss each one in turn in detail.

Big Idea 1. Distributions describe variability in data.

Essential Understanding 1*a*. Graphs and tables are useful for displaying distributions of categorical data.

Essential Understanding 1*b*. Numerical summaries of categorical data are useful for describing particular features of a distribution.

Essential Understanding 1*c*. Graphs and tables are useful for displaying distributions of quantitative data.

Essential Understanding 1*d*. Numerical summaries of quantitative data are useful for measuring the center of a distribution.

Essential Understanding 1*e*. Numerical summaries of quantitative data are useful for measuring the amount of variability within a distribution.

Essential Understanding 1*f*. Graphs and tables based on grouped data are useful for displaying distributions of quantitative data.

Essential Understanding 1*g*. The shape of a distribution influences which summary measure is most appropriate for describing the center of a distribution for quantitative data.

Essential Understanding 1*h*. Graphs and tables based on a division of the ordered data into equal-sized groups are useful for displaying distributions of quantitative data.

Essential Understanding 1*i*. Some numerical summaries of quantitative data are more resistant than others to extreme data values, called *outliers*.

Big Idea 2. Statistics can be used to compare two or more groups of data.

Essential Understanding 2*a*. The focus of comparisons between two or more groups of data is on similarities and differences between the distributions.

Essential Understanding 2*b*. The amount of separation between two or more distributions is related to the amount of variability within them.

Big Idea 3. Bivariate distributions describe patterns or trends in the covariability in data on two variables.

Essential Understanding 3*a*. Graphs and tables are useful for displaying bivariate distributions of data on two categorical variables.

Essential Understanding 3*b*. Conditional relative frequency distributions are useful for establishing an association between two categorical variables.

Essential Understanding 3*c*. Graphs and tables are useful for displaying bivariate distributions of data on two quantitative variables.

Essential Understanding 3*d*. A correlation coefficient is a numerical summary of bivariate data that measures the strength of the relationship between two variables.

Essential Understanding 3e. When the trend in bivariate data on two quantitative variables is generally linear, a centrally located line can be useful for making predictions.

Big Idea 4. Inferential statistics uses data in a sample selected from a population to describe features of the population.

Essential Understanding 4*a*. The sampling distribution of a statistic describes the sample-to-sample variability in values of the statistic from multiple samples of the same size selected from the same population.

Essential Understanding 4*b*. Selecting a simple random sample from a population is a fair way to select a sample.

Essential Understanding 4c. The predictable pattern for the sampling distribution of a statistic based on random sampling provides a way for making inferences about the population.