# 1

# THE NATURE OF STATISTICS



## 1.1 STATISTICS DEFINED

Broadly defined, *statistics* involves the theory and methods of

collecting,
organizing,
presenting,
analyzing, and
interpreting

data so as to determine their essential characteristics. While some discussion will be devoted to the collection, organization, and presentation of data, we shall, for the most part, concentrate on the analysis of data and the interpretation of the results of our analysis.

How should the notion of *data* be viewed? It can be thought of as simply consisting of "information" that can take a variety of forms. For example, data can be *numerical* (test scores, weights, lengths, elapsed time in minutes, etc.) or *non-numerical* (such as an attribute involving color or texture or a category depicting the sex of an individual or their political affiliation, if any, etc.) (See Section 1.4 of this chapter for a more detailed discussion of data forms or varieties.)

*Statistical Inference: A Short Course*, First Edition. Michael J. Panik.

Two major types of statistics will be recognized: (1) *descriptive*; and (2) *inductive*[1] or *inferential.*

*Descriptive Statistics*: Deals with summarizing data. Our goal here is to arrange data in a readable form. To this end, we can construct tables, charts, and graphs; we can also calculate percentages, rates of change, and so on. We simply offer a picture of "what is" or "what has transpired."

*Inductive Statistics*: Employs the notion of *statistical inference*, that is, inferring something about the entire data set from an examination of only a portion of the data set. How is this inferential process carried out? Through *sampling*—a representative group of items is subject to study and the conclusions derived therefrom are assumed to characterize the entire data set. Keep in mind, however, that since we are only sampling and not undertaking an exhaustive census of the entire data set, some "margin of error" associated with our inference will most assuredly emerge. Hence, our sample result must be accompanied by a measure of the uncertainty of the inference made. Questions such as "How reliable is our result?" or, "What is the level of confidence associated with our result?" must be addressed before presenting our findings. This is why inferential statistics is often referred to as "decision making under uncertainty." Clearly inferential statistics enables us to go beyond a purely descriptive treatment of data—it enables us to make estimates, forecasts or predictions, and generalizations.

In sum, if we want to only summarize or present data or just catalog facts then descriptive techniques are called for. But if we want to make inferences about the entire data set on the basis of sample information or, more generally, make decisions in the face of uncertainty then the use of inductive or inferential techniques is warranted.

## 1.2 THE POPULATION AND THE SAMPLE

The concept of the "entire data set" alluded to above will be called the *population*; it is the group to be studied. (Remember that "population" does not refer exclusively to "people;" it can be a group of states, countries, cities, registered democrats, cars in a parking lot, students at a particular academic institution, and so on.) We shall let $N$ denote the population size or the number of elements in the population.

Each separate characteristic of an element in the population will be represented by a variable (usually denoted as $X$). We may think of a *variable* as describing any qualitative or quantitative aspect of a member of the population. A *qualitative* variable has values that are only "observed." Here a characteristic pertains to some

[1] *Induction* is a process of reasoning from the specific to the general.

attribute (such as color) or category (male or female). A *quantitative* variable will be classified as either *discrete* (it takes on a finite or countable number of values) or *continuous* (it assumes an infinite or uncountable number of values). Hence, discrete values are "counted;" continuous values are "measured." For instance, a discrete variable might be the number of blue cars in a parking lot, the number of shoppers passing through a supermarket check-out counter over a 15 min time interval, or the number of sophomores in a college-level statistics course. A continuous variable can describe weight, length, the amount of water passing through a culvert during a thunderstorm, elapsed time in a race, and so on.

While a population can consist of all conceivable observations on some variable $X$, we may view a *sample* as a subset of the population. The sample size will be denoted as $n$, with $n < N$. It was mentioned above that, in order to make a legitimate inference about a population, a *representative sample* was needed. Think of a representative sample as a "typical" sample—it should adequately reflect the attributes or characteristics of the population.

## 1.3 SELECTING A SAMPLE FROM A POPULATION

While there are many different ways of constructing a sampling plan, our attention will be focused on the notion of simple random sampling. Specifically, a sample of size $n$ drawn from a population of size $N$ is obtained via *simple random sampling* if every possible sample of size $n$ has an equal chance of being selected. A sample obtained in this fashion is then termed a *simple random sample*; each element in the population has the same chance of being included in a simple random sample.

Before any sampling is actually undertaken, a list of items (called the *sampling frame*) in the population is formed and thus serves as the formal source of the sample, with the individual items listed on the frame termed *elementary sampling units*. So, given the sampling frame, the actual process of random sample selection will be accomplished *without replacement*, that is, once an item from the population has been selected for inclusion in the sample, it is not eligible for selection again—it is not returned to the population pool (it is, so to speak, "crossed off" the frame) and consequently cannot be chosen, say, a second time as the simple random sampling process commences. (Under *sampling with replacement*, the item chosen is returned to the population before the next selection is made.)

Will the process of random sampling guarantee that a representative sample will be acquired? The answer is, "probably." That is, while randomization does not absolutely guarantee representativeness (since random sampling gives the same chance of selection to every sample—representative ones as well as nonrepresentative ones), we are highly likely but not certain to get a representative sample. Then why all the fuss about random sampling? The answer to this question hinges upon the fact that it is possible to make erroneous inferences from sample data. (After all, we are not examining the entire population.) Under simple random sampling, we can validly apply the rules of probability theory to calculate the chances or magnitudes of such

errors; and their rates enable us to assess the reliability of, or form a degree of confidence in, our inferences about the population.

Let us recognize two basic types of errors that can creep into our data analysis. The first is *sampling error*, which is reflective of the inherent natural variation between samples (since different samples possess different sample values); it arises because sampling gives incomplete information about a population. This type of error is inescapable—it is always present. If one engages in sampling then sampling error is a fact of life. The other variety of error is *nonsampling* error—human or mechanical factors tend to distort the observed values. Nonsampling error can be controlled since it arises essentially from unsound experimental techniques or from obtaining and recording information. Examples of nonsampling error can range from using poorly calibrated or inadequate measuring devices to inaccurate responses (or nonresponses) to questions on a survey form. In fact, even poorly worded questions can lead to such errors. And if preference is given to selecting some observations over others so that, for example, the underrepresentation of some group of individuals or items occurs, then a *biased* sample results.

## 1.4 MEASUREMENT SCALES

We previously referred to *data*[2] as "information," that is, as a collection of facts, values, or observations. Suppose then that our data set consists of observations that can be "measured" (e.g., classified, ordered, or quantified). At what level does the measurement take place? In particular, what are the "forms" in which data are found or the "scales" on which data are measured? These scales, offered in terms of increasing information content, are classified as nominal, ordinal, interval, and ratio.

1. *Nominal Scale*: Nominal should be associated with the word "name" since this scale identifies categories. Observations on a nominal scale possess neither numerical value nor order. A variable whose values appear on a nominal scale is termed qualitative or categorical. For example, a variable $X$ depicting the sex of an individual (male or female) is nominal in nature as are variables depicting religion, political affiliation, occupation, marital status, color, and so on. Clearly, nominal values cannot be ranked or ordered—all items are treated equally. The only valid operations for variables treated on a nominal scale are the determination of "$=$" or "$\neq$." For nominal data, any statistical analysis is limited and usually relegated to the calculation of percentages.
2. *Ordinal Scale*: (think of the word "order") Includes all properties of the nominal scale with the additional property that the observations can be ranked from the "least important" to the "most important." For instance, hierarchical

[2] "Data" is a plural noun; "datum" is the singular of data.

organizations within which some members are more important or ranked higher than others have observations that are considered to be ordinal since a "pecking order" can be established. For example, military organizations exhibit a well-defined hierarchy (although it is "better" to be a colonel than a private, the ranking does not indicate "how much better"). Other examples are as follows:

| Performance Ratings | Corporate Hierarchy |
|---|---|
| Excellent | President |
| Very good | Senior vice president |
| Good | Vice president |
| Fair | Assistant vice president |
| Poor | Senior manager |
| | Manager |

The only valid operations for ordinally scaled variables are "$=, \neq, <, >$." Both nominal and ordinal scales are *nonmetric scales* since differences among their values are meaningless.

3. *Interval Scale*: Includes all the properties of the ordinal scale with the additional property that the distance between observations is meaningful; the numbers assigned to the observations indicate order and possess the property that the difference between any two consecutive values is the same as the difference between any other two consecutive values. Hence, the difference $3 - 2 = 1$ has the same meaning as $5 - 4 = 1$. While an interval scale has a zero point, its location may be arbitrary so that ratios of interval scale values have no meaning. For instance, 0°C does not imply the absence of heat (it is simply the temperature at which water freezes); or 60°C is not twice as hot as 30°C. Also, a score of zero on a standardized test does not imply a lack of knowledge; and a student with a score of 400 is not four times as smart as a student who scored 100. The operations for handling variables measured on an interval scale are "$=, \neq, <, >, +, -$."
4. *Ratio Scale*: Includes all the properties of the interval scale with the added property that ratios of observations are meaningful. This is because "absolute zero is uniquely defined." In this regard, if a variable $X$ is measured in dollars ($), then $0 represents the "absence of monetary value;" and a price of $20 is twice as costly as a price of $10 (the ratio is $2/1 = 2$). Other examples of ratio scale measurements are as follows: weight, height, age, GPA, income, and so on. Valid operations for variables measured on a ratio scale are "$=, \neq, <, >, +, -, \times, \div$."

Both interval and ratio scales are said to be *metric scales* since differences between values measured on these scales are meaningful; and variables measured on these scales are said to be *quantitative variables*.

It should be evident from the preceding discussion that any variable measured on one scale automatically satisfies all the properties of a less informative scale.

**TABLE 1.1 Characteristics of the Residence at 401 Elm Street**

| Characteristic | Variable | Observation Values |
|---|---|---|
| Number of families | $X_1$ | One family (1) or two family (2) |
| Attached garage | $X_2$ | Yes (1) or no (0) |
| Number of rooms | $X_3$ | 6 |
| Number of bathrooms | $X_4$ | 2 |
| Square feet of living space | $X_5$ | 2100 |
| Assessed value | $X_6$ | $230,500 |
| Year constructed | $X_7$ | 1987 |
| Lot size (square feet) | $X_8$ | 2400 |

### Example 1.1

Suppose our objective is to study the residential housing stock of a particular city. Suppose further that our inquiry is to be limited to one- and two-family dwellings. These categories of dwellings make up the *target population*—the population about which information is desired. How do we obtain data on these housing units? Should we simply stroll around the city looking for one- and two-family housing units? Obviously this would be grossly inefficient. Instead, we will consult the City Directory. This directory is the *sampled population* (or sampling frame)—the population from which the sample is actually obtained. Now, if the City Directory is kept up to date then we have a *valid sample*—the target and sampled populations have similar characteristics.

Let the individual residences constitute the elementary sampling units, with an *observation* taken to be a particular data point or value of some characteristic of interest. We shall let each such characteristic be represented by a separate variable, and the value of the variable is an observation of the characteristic.

Assuming that we have settled on a way of actually extracting a random sample from the directory, suppose that one of the elementary sampling units is the residence located at 401 Elm Street. Let us consider some of its important characteristics (Table 1.1).

Note that $X_1$ and $X_2$ are qualitative or nonmetric variables measured on a nominal scale while variables $X_3, \ldots, X_8$ are quantitative or metric variables measured on a ratio scale.

## 1.5 LET US ADD

Quite often throughout this text the reader will be asked to total or form the sum of the values appearing in a variety of data sets. We have a special notation for the operation

of addition. We will let the Greek capital sigma or $\Sigma$ serve as our "summation sign." Specifically, for a variable $X$ with values $X_1, X_2, \ldots, X_n$,

$$X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i. \tag{1.1}$$

Here the right-hand side of this expression reads "the sum of all observations $X_i$ as $i$ goes from 1 to $n$." In this regard, $\Sigma$ is termed an *operator*—it operates only on those items having an $i$ index, and the operation is addition. When it is to be understood that we are to add over all $i$ values, then Equation (1.1) can be rewritten simply as $X_1 + X_2 + \cdots + X_n = \Sigma X_i$.

## EXERCISES

**1** Are the following variables qualitative or quantitative?

**a.** Color
**b.** Gender
**c.** Zip code
**d.** Temperature
**e.** Your cell phone number
**f.** Posted speed limit
**g.** Life expectancy
**h.** Size of your family
**i.** Number on a uniform
**j.** Marital status
**k.** Drivers license number
**l.** Cost to fill up your car's gas tank

**2** Are the following variables discrete or continuous?

**a.** Number of trucks parked at a truck stop
**b.** Number of tails obtained in three flips of a fair coin
**c.** Time taken to walk up a flight of stairs
**d.** The height attained by a high jumper
**e.** Life expectancy
**f.** Number of runs scored in a baseball game
**g.** Length of your favorite song
**h.** Weight loss after dieting for a month

**3** Are the following variables nominal or ordinal?

**a.** Gender
**b.** Brand of a loaf of bread
**c.** Response to a customer satisfaction survey: poor, fair, good, or excellent
**d.** Letter grade in a college course
**e.** Faculty rank at a local college

**4** Are the following variables all measured on a ratio scale?

**a.** Cost of a new pair of shoes
**b.** A day's wages for a laborer
**c.** Your house number
**d.** Porridge in the pot 9 days old
**e.** Your shoe size
**f.** An IQ score of 130