# I. POPULATIONS, VARIABLES, and DATA

### **Populations and Samples:**

To a statistician, the *population* is the set or collection under investigation. Individual members of the population are not usually of interest. Rather, investigators try to infer with some degree of confidence the general features of the population.

# **Examples:**

- 1. The population of students currently enrolled at a certain university.
- 2. The population of registered voters in a certain Congressional district.
- 3. The population of large-mouthed bass in a certain lake.
- 4. The population of all decay times of a certain species of radioactive nucleus.

From these examples we see that many populations are so large that it is impossible to examine each of their individual members. In fact, the populations in A - C change over time. Population D is so large that we might well consider it to be infinite. Nevertheless, the techniques of statistical inference allow us to draw conclusions about these populations from observations made on a smaller subset of the population. This smaller subset is called a *sample* from the population. We shall say more about samples later.

### Variables:

A *population variable* is a descriptive number or label associated with each member of a population. The values of a population variable are the various numbers (or labels) that occur as we consider all the members of the population. Values of variables that have been recorded for a population or a sample from a population constitute *data*.

# Example 1.1:

Consider the population of all students currently enrolled at this university. The university maintains a database with values of many variables recorded for each student. Among them are "Gender", "Academic Classification", "Earned Hours", "Grade Points", "Zip Code", "Social Security Number", "Age", and others. Other variables for which the university has no data might also be defined, for example, "Total Income in 2003" and "Number of Cousins".

# **Types of Data:**

Statisticians sometimes classify variables or data into a hierarchy of types. The different types correspond to the kinds of meaningful operations that can be performed on data and the ways it can best be represented or displayed. The list below progresses from very little structure to a considerable degree of structure.

- A. *Nominal* variables are variables whose values are labels. The order of the labels may have no special significance. In the example above, "Gender" is a nominal variable whose values are "M" and "F". Nominal variables are also called *categorical variables* or *factors*. Nominal variables may be represented by numbers. If so, these numbers are used merely as labels and are not subjected to arithmetic operations. "Zip Code" is a nominal variable whose values are represented by numbers.
- B. *Ordinal* variables are variables whose values have a natural order. If they are represented as numbers, the order of the numerical values should reflect the natural ordering. Responses to an item on a questionnaire, ranging from "Strongly Disagree" to "Strongly Agree" are ordinal data. The letter grades assigned by a teacher to a class of students are ordinal data. Ordinal variables such as these that have only a small number of values are sometimes called *ordered categories* or *ordered factors*.
- C. *Interval* variables have values represented by numbers. Both the order of the values and the difference between any two values are meaningful. To represent an interval variable, a definite unit of measurement is used (e.g., meters, seconds, pounds, credit hours, etc.) The numerical values of an interval variable may be in reference to a somewhat arbitrary zero point. For example, the locations of automobile accidents on Interstate 10 can be given in miles west from downtown San Antonio or from downtown Houston. No matter which reference point we use, it makes sense to say that one accident occurred 20 miles west of another accident. Interval variables might have either positive or negative values. Interval variables might also be considered ordinal, but in general interval variables have more structure than ordinal variables.
- D. *Ratio* variables have values that are positive numbers on a scale with a unit of measurement and a natural zero point. Both differences and ratios of values are meaningful. Examples are "Age", "Weight", and "Annual Income". Ratio variables might also be considered interval variables.

It should be obvious that there will be many variables that are not easily classified according to this scheme. If the decision is really difficult, then it probably doesn't matter very much exactly how a variable is classified. Some authors simply classify variables as qualitative or quantitative without making finer distinctions.

**Exercise:** Classify the variables mentioned in the example of the student database. Think of some other variables that might be included and classify them.

**Exercise:** Look at some of the data sets provided with this course. For each one, describe the population if you can. Identify the variables and classify them as nominal, ordinal, interval, or ratio.