What is statistics?

Definition

Statistics *Statistics* is the science of collecting, analyzing, presenting, and interpreting data, as well as of making decisions based on such analyses.

1.1.1 What Is Statistics?

The word **statistics** has two meanings. In the more common usage, *statistics* refers to numerical facts. The numbers that represent the income of a family, the age of a student, the percentage of passes completed by the quarterback of a football team, and the starting salary of a typical college graduate are examples of statistics in this sense of the word. A 1988 article in *U.S. News & World Report* mentioned that "Statistics are an American obsession."[2] During the 1988 baseball World Series between the Los Angeles Dodgers and the Oakland A's, the then NBC commentator Joe Garagiola reported to the viewers numerical facts about the players' performances. In response, fellow commentator Vin Scully said, "I love it when you talk statistics." In these examples, the word *statistics* refers to numbers.

Nature of Statistics

1. **Data Collection and Organization:**
   - Statistics involves the collection, organization, and presentation of data. This could be in the form of surveys, experiments, or observations.
2. **Statistical Analysis:**
   - Once data is collected, statistics helps in analyzing and interpreting it. This includes the use of various statistical methods and techniques to draw meaningful conclusions.
3. **Variability and Uncertainty:**
   - Statistics deals with variability and uncertainty in data. It recognizes that not all observations are identical and seeks to understand and quantify this variability.

4. **Decision Making under Uncertainty:**
   - One of the primary goals of statistics is to aid decision-making in the face of uncertainty. It provides tools for making informed decisions based on available data.
5. **Generalization:**
   - Statistics allows for generalization from a sample to a population. By studying a subset (sample) of a larger group (population), statisticians make inferences about the entire population.

Elements of Basic Statistics

Basic statistics involves the collection, analysis, interpretation, presentation, and organization of data. Here are some fundamental elements of basic statistics:

Descriptive Statistics:

Definition: Descriptive statistics summarize and describe the main features of a dataset.
Example: Calculating the mean (average), median, and mode of a set of exam scores to understand the central tendency.

Inferential Statistics:

Definition: Inferential statistics involve making inferences or predictions about a population based on a sample of data.
Example: Using the average score of a sample to estimate the average score of an entire class.

3. **Biostatistics:**
    - Biostatistics applies statistical methods to biological and health-related data. It is used in clinical trials, epidemiology, and other areas of medical research to draw conclusions about the effectiveness of treatments or the prevalence of diseases.
4. **Econometrics:**
    - Econometrics applies statistical methods to economic data. It helps economists analyze economic relationships, test hypotheses, and forecast future economic trends.

5. **Quality Control and Six Sigma:**
    - Statistics plays a crucial role in quality control processes in manufacturing and service industries. Six Sigma, for example, relies heavily on statistical methods to improve process efficiency and reduce defects.
6. **Social Sciences:**
    - Statistics is extensively used in social sciences for research in sociology, psychology, political science, and more. Surveys and experiments in these fields often involve statistical analysis to draw meaningful conclusions.

7. **Environmental Statistics:**
    - Environmental scientists use statistics to analyze and interpret data related to environmental issues such as pollution levels, climate change, and biodiversity.

In summary, the nature of statistics encompasses data collection, analysis, and decision-making under uncertainty, while its scope extends to various fields, making it a crucial tool for making informed decisions in a wide range of disciplines.

Data Analysis in Research and Development:

- Statistical analysis is fundamental in research and development processes. It helps engineers make informed decisions, assess the significance of experimental results, and identify trends or patterns in data.

Simulation and Modeling:

- Engineers often use statistical methods to simulate and model complex systems. Monte Carlo simulations, for example, rely on statistical sampling techniques to model the behavior of systems under various conditions.

Risk Analysis:

- Engineers use statistical tools for risk analysis and decision-making. This is crucial in evaluating potential risks associated with design choices, project timelines, and budget estimates.

1.2 Population Versus Sample

We will encounter the terms *population* and *sample* on almost every page of this text.[3] Consequently, understanding the meaning of each of these two terms and the difference between them is crucial.

Suppose a statistician is interested in knowing the following:

1. The percentage of all voters in a city who will vote for a particular candidate in an election
2. Last year's gross sales of all companies in New York City
3. The prices of all houses in California

In these examples, the statistician is interested in *all* voters in a city, *all* companies in New York City, and *all* houses in California. Each of these groups is called the population for the respective example. In statistics, a population does not necessarily mean a collection of people. It can, in fact, be a collection of people or of any kind of item such as houses, books, television sets, or cars. The population of interest is usually called the **target population**.

Definition

Population or Target Population A *population* consists of all elements—individuals, items, or objects—whose characteristics are being studied. The population that is being studied is also called the *target population*.

Most of the time, decisions are made based on portions of populations. For example, the election polls conducted in the United States to estimate the percentages of voters who favor various candidates in any presidential election are based on only a few hundred or a few thousand voters selected from across the country. In this case, the population consists of all registered voters in the United States. The sample is made up of a few hundred or few thousand voters who are included in an opinion poll. Thus, the collection of a number of elements selected from a population is called a **sample**. Figure 1.1 illustrates the selection of a sample from a population.

Definition

Sample A portion of the population selected for study is referred to as a *sample*.

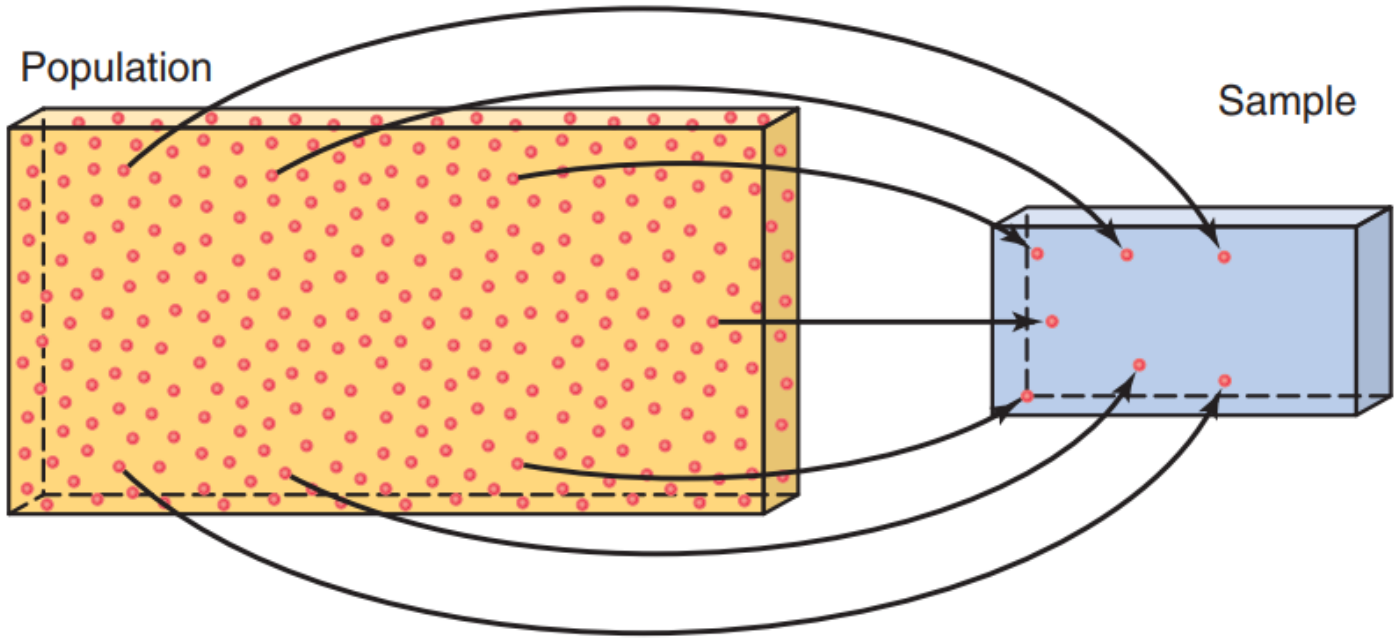

Figure 1.1 Population and sample.

Definition

Representative Sample A sample that represents the characteristics of the population as closely as possible is called a *representative sample.*

A sample may be random or nonrandom. In a **random sample**, each element of the population has a chance of being included in the sample. However, in a nonrandom sample this may not be the case.

Definition

Random Sample A sample drawn in such a way that each element of the population has a chance of being selected is called a *random sample.* If all samples of the same size selected from a population have the same chance of being selected, we call it **simple random sampling**. Such a sample is called a **simple random sample**.

One way to select a random sample is by lottery or draw. For example, if we are to select 5 students from a class of 50, we write each of the 50 names on a separate piece of paper. Then we place all 50 slips in a box and mix them thoroughly. Finally, we randomly draw 5 slips from the box. The 5 names drawn give a random sample. On the other hand, if we arrange all 50 names alphabetically and then select the first 5 names on the list, it is a nonrandom sample because the students listed 6th to 50th have no chance of being included in the sample.

A sample may be selected with or without replacement. In sampling **with replacement**, each time we select an element from the population, we put it back in the population before we

1.3 Basic Terms

It is very important to understand the meaning of some basic terms that will be used frequently in this text. This section explains the meaning of an element (or member), a variable, an observation, and a data set. An element and a data set were briefly defined in Section 1.1. This section defines these terms formally and illustrates them with the help of an example.

Table 1.1 gives information on the total revenues (in millions of U.S. dollars) for the year 2010 of the top six revenue-earning companies in the world. We can call this group of companies a sample of six companies. (Note that it is not a random sample.) Each company listed in this table is called an **element** or a **member** of the sample. Table 1.1 contains information on six elements. Note that elements are also called *observational units*.

Definition

Element or Member An *element* or *member* of a sample or population is a specific subject or object (for example, a person, firm, item, state, or country) about which the information is collected.

Table 1.1 Total Revenues for 2010 of Six Companies

Company	2010 Total Revenue (millions of dollars)
Wal-Mart Stores	421,849
Royal Dutch Shell	378,152
Exxon Mobil	354,674
BP	308,928
Sinopec Group	273,422
China National Petroleum	240,192

Source: Fortune Magazine, July 25, 2011.

An element or a member → Exxon Mobil

← Variable

354,674 ← An observation or measurement

What is Data?

In statistics, "data" refers to any collection of information or observations that are systematically recorded and can be analyzed. Data can take various forms, including numbers, words, measurements, and more. There are two main types of data: qualitative (categorical) and quantitative (numerical).

1. **Qualitative Data:**

This type of data represents categories or labels and cannot be measured in numerical terms. Examples include:
Colors: red, blue, green
Marital status: single, married, divorced
Types of fruit: apple, orange, banana

2. Quantitative Data:

This type of data consists of numerical measurements or counts. It can be further divided into two subtypes: discrete and continuous.

Discrete Data:

Consists of distinct, separate values that can be counted. Examples include:

Number of students in a class: 30, 45, 60 (whole numbers)

Number of cars in a parking lot: 10, 20, 30 (whole numbers)

Continuous Data:

Represents measurements that can take any value within a given range. Examples include:

Height of individuals: 160.5 cm, 175.2 cm, 182.0 cm (any real number)

Weight of objects: 3.2 kg, 5.7 kg, 10.1 kg (any real number)

Variable

Definition

Variable A *variable* is a characteristic under study that assumes different values for different elements. In contrast to a variable, the value of a *constant* is fixed.

A few other examples of variables are household incomes, the number of houses built in a city per month during the past year, the makes of cars owned by people, the gross profits of companies, and the number of insurance policies sold by a salesperson per day during the past month.

In general, a variable assumes different values for different elements, as do the 2010 revenues for the six companies in Table 1.1. For some elements in a data set, however, the values of the variable may be the same. For example, if we collect information on incomes of households, these households are expected to have different incomes, although some of them may have the same income.

A variable is often denoted by x, y, or z. For instance, in Table 1.1, the 2010 revenue for companies may be denoted by any one of these letters. Starting with Section 1.7, we will begin to use these letters to denote variables.

Each of the values representing the 2010 revenues of the six companies in Table 1.1 is called an **observation** or **measurement**.

Definition

Observation or Measurement The value of a variable for an element is called an *observation* or *measurement.*

From Table 1.1, the 2010 revenue of Exxon Mobil was $354,674 million. The value $354,674 million is an observation or a measurement. Table 1.1 contains six observations, one for each of the six companies.

The information given in Table 1.1 on the 2010 revenues of companies is called the **data** or a **data set**.

Definition

Data Set A *data set* is a collection of observations on one or more variables.

Other examples of data sets are a list of the prices of 25 recently sold homes, test scores of 15 students, opinions of 100 voters, and ages of all employees of a company.

1.4 Types of Variables

In Section 1.3, we learned that a variable is a characteristic under investigation that assumes different values for different elements. Family income, height of a person, gross sales of a company, price of a college textbook, make of the car owned by a family, number of accidents, and status (freshman, sophomore, junior, or senior) of a student enrolled at a university are examples of variables.

A variable may be classified as quantitative or qualitative. These two types of variables are explained next.

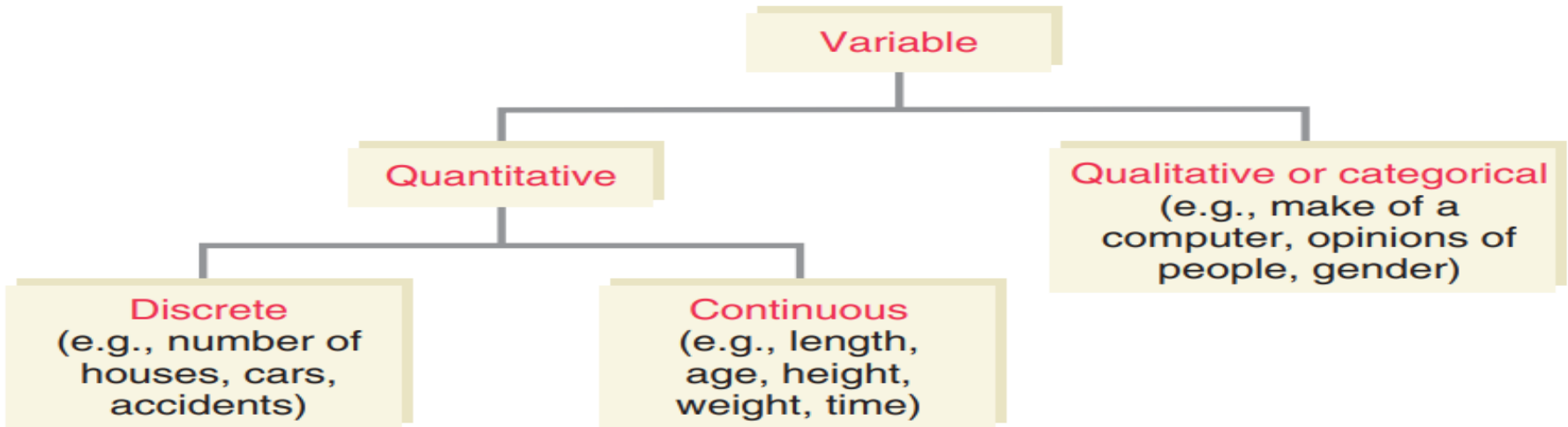

Figure 1.2 Types of variables.

1.4.1 Quantitative Variables

Some variables (such as the price of a home) can be measured numerically, whereas others (such as hair color) cannot. The first is an example of a **quantitative variable** and the second that of a qualitative variable.

Definition

Quantitative Variable A variable that can be measured numerically is called a *quantitative variable.* The data collected on a quantitative variable are called *quantitative data.*

Incomes, heights, gross sales, prices of homes, number of cars owned, and number of accidents are examples of quantitative variables because each of them can be expressed numerically.

Discrete Variables

The values that a certain quantitative variable can assume may be countable or noncountable. For example, we can count the number of cars owned by a family, but we cannot count the height of a family member. A variable that assumes countable values is called a **discrete variable**. Note that there are no possible intermediate values between consecutive values of a discrete variable.

Definition

Discrete Variable A variable whose values are countable is called a *discrete variable*. In other words, a discrete variable can assume only certain values with no intermediate values.

For example, the number of cars sold on any given day at a car dealership is a discrete variable because the number of cars sold must be 0, 1, 2, 3, . . . and we can count it. The number of cars sold cannot be between 0 and 1, or between 1 and 2. Other examples of discrete variables are the number of people visiting a bank on any day, the number of cars in a parking lot, the number of cattle owned by a farmer, and the number of students in a class.

Continuous Variables

Some variables cannot be counted, and they can assume any numerical value between two numbers. Such variables are called **continuous variables**.

Definition

Continuous Variable A variable that can assume any numerical value over a certain interval or intervals is called a *continuous variable.*

The time taken to complete an examination is an example of a continuous variable because it can assume any value, let us say, between 30 and 60 minutes. The time taken may be 42.6 minutes, 42.67 minutes, or 42.674 minutes. (Theoretically, we can measure time as precisely as we want.) Similarly, the height of a person can be measured to the tenth of an inch or to the hundredth of an inch. However, neither time nor height can be counted in a discrete fashion. Other examples of continuous variables are weights of people, amount of soda in a 12-ounce can (note that a can does not contain exactly 12 ounces of soda), and yield of potatoes (in pounds) per acre. Note that any variable that involves money and can assume a large number of values is typically treated as a continuous variable.

1.4.2 Qualitative or Categorical Variables

Variables that cannot be measured numerically but can be divided into different categories are called **qualitative** or **categorical variables**.

Definition

Qualitative or Categorical Variable A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a *qualitative* or *categorical variable.* The data collected on such a variable are called *qualitative data.*

For example, the status of an undergraduate college student is a qualitative variable because a student can fall into any one of four categories: freshman, sophomore, junior, or senior. Other

THANK YOU FOR
YOUR
CONCENTRATIONS