

## 1.2: Definitions of Statistics, Probability, and Key Terms

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives. **Data** represent all the pieces of information or observations collected on characteristics of our interest (actual values of the variable). They may be numbers or they may be words. **Datum** is a single value. Data can come from a population or a sample.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

### Probability

**Probability** is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is  $\frac{1}{2}$  or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction  $\frac{996}{2000}$  is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a successful surgery or medical treatment. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

### Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter, in this case the mean. A **parameter** is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a

**representative sample.** We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, or random variable, usually denoted by capital letters such as  $X$  and  $Y$ , is a characteristic or measurement that can be determined for each member of a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let  $X$  equal the number of points earned by one math student at the end of a term, then  $X$  is a numerical variable. If we let  $Y$  be a person's party affiliation, then some examples of  $Y$  include Republican, Democrat, and Independent.  $Y$  is a categorical variable. We could do some math with values of  $X$  (calculate the average number of points earned, for example), but it makes no sense to do math with values of  $Y$  (calculating an average party affiliation makes no sense).

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is  $\frac{22}{40}$  and the proportion of women students is  $\frac{18}{40}$ . Mean and proportion are discussed in more detail in later chapters.

#### Note

The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

#### Example 1.2.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

#### Answer

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term: the population mean.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let  $X$  = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are \$150, \$200, and \$225.

#### Exercise 1.2.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

#### Example 1.2.2

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. Population \_\_\_\_ 2. Statistic \_\_\_\_ 3. Parameter \_\_\_\_ 4. Sample \_\_\_\_ 5. Variable \_\_\_\_ 6. Data \_\_\_\_

- a. all students who attended the college last year
- b. the cumulative GPA of one student who graduated from the college last year
- c. 3.65, 2.80, 1.50, 3.90
- d. a group of students who graduated from the college last year, randomly selected
- e. the average cumulative GPA of students who graduated from the college last year
- f. all students who graduated from the college last year
- g. the average cumulative GPA of students in the study who graduated from the college last year

**Answer**

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

**Example 1.2.3**

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used. Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

**Answer**

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable**  $X$  = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

**Example 1.2.4**

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

**Answer**

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable**  $X$  = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

---

This page titled [1.2: Definitions of Statistics, Probability, and Key Terms](#) is shared under a [CC BY](#) license and was authored, remixed, and/or curated by .