

## CHAPTER 1 Picturing Distributions with Graphs



mark@rocketclips.com/Depositphotos

### IN THIS CHAPTER WE COVER...

- [Individuals and variables](#)
- [Identifying categorical and quantitative variables](#)
- [Categorical variables: pie charts and bar graphs](#)
- [Quantitative variables: histograms](#)
- [Interpreting histograms](#)
- [Quantitative variables: dotplots](#)
- [Time plots](#)
- [Discussion: \(Mis\)adventures in data entry](#)

**S**tatistics is the science of data. Data collection has increased tremendously in the new millennium, and this trend shows no signs of slowing down. In 2003, the Human Genome Project uncovered, after 13 years of international cooperation and billions of dollars, the complete sequence of the 3 billion DNA bases of the human genome. A decade later, individuals can have their own genome sequenced within weeks for just thousands of dollars. In the United States, the Census Bureau collects extensive data on the nation from numerous surveys, such as the yearly National Health Interview Survey, which gathers health and socioeconomic information on each member of 40,000 households. Gallup, a private polling firm, surveys roughly 1000 U.S. adults per day for its Gallup-Healthways Well-Being Index covering a whole range of physical and mental health conditions. Smartphones have become so ubiquitous that they are used to monitor ongoing epidemics worldwide. Satellites provide detailed daily records of climatic conditions at the planetary level. Millions of individuals now routinely use apps and wearables to collect and store a wealth of health data, which are then carefully examined by private individuals and large corporations alike. “Big data” is an exploding phenomenon that has brought about a renewed interest in data exploration. While some

of the techniques used with extremely large and complex data sets are very specific to their unique needs, the fundamental principles are essentially the same as those we will study in this and the following few chapters.

USFCA  
ommah@usfca.edu

## INDIVIDUALS AND VARIABLES

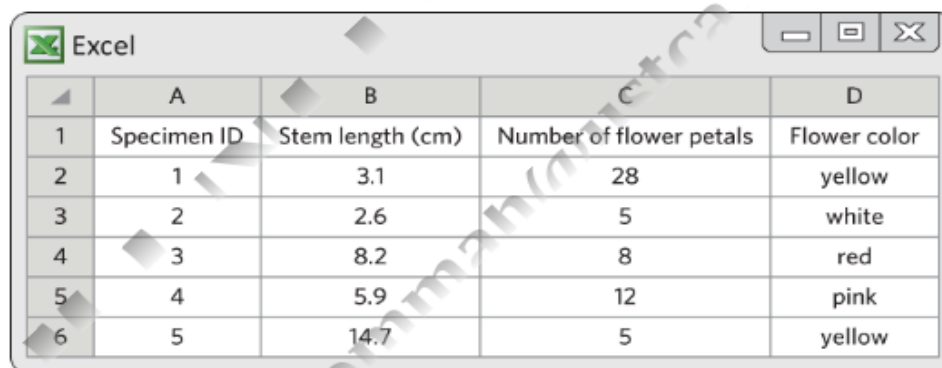
Any set of data contains information about some group of *individuals*. The information is organized in *variables*. The techniques of descriptive statistics covered in [Part I](#) of this book apply equally to data sets obtained from a given **population** (the entire group of individuals about which we want information) and to data sets collected from only those individuals in a smaller **sample**. The distinction between a population and a sample is an important one for inference, and we will address it in more detail in [Chapter 6](#).

### INDIVIDUALS AND VARIABLES

**Individuals** are the objects (or units) described by a set of data. Individuals may be people, but they may also be animals, plants, or things.

A **variable** is any one characteristic of an individual. A variable can take different values for different individuals.

A botanist's plant database, for example, includes data about various aspects of the plants examined. The plants are the individuals described by the database. For each individual (each plant), the data contain the values of variables such as stem length, number of flower petals, and flower color. An example of what such a database might look like is shown in [Figure 1.1](#), with each row dedicated to one individual and each column to a variable. **Spreadsheet** programs with rows and columns readily available are commonly used to store data and do simple calculations.



	A	B	C	D
1	Specimen ID	Stem length (cm)	Number of flower petals	Flower color
2	1	3.1	28	yellow
3	2	2.6	5	white
4	3	8.2	8	red
5	4	5.9	12	pink
6	5	14.7	5	yellow

**Figure 1.1**  
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

**FIGURE 1.1** Example of a botanical database displayed in a spreadsheet. Each column includes data about a different variable. Each row represents data for one plant specimen.

Note that *sometimes the individuals represented in a data set are actually groups*, such as a set of countries, states, or counties, or even a set of ant colonies. For instance, every year, the U.S. Centers for Disease Control and Prevention (CDC) reports the percent of obese individuals in each state. In this data set, the 50 states are the individuals studied and the variable recorded for each individual is the percent of the population in that state who are obese.



## INDIVIDUALS AND VARIABLES

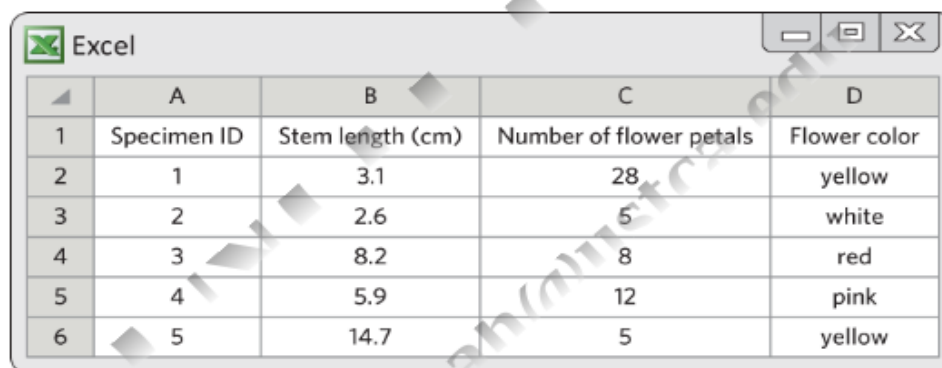
Any set of data contains information about some group of *individuals*. The information is organized in *variables*. The techniques of descriptive statistics covered in [Part I](#) of this book apply equally to data sets obtained from a given **population** (the entire group of individuals about which we want information) and to data sets collected from only those individuals in a smaller **sample**. The distinction between a population and a sample is an important one for inference, and we will address it in more detail in [Chapter 6](#).

### INDIVIDUALS AND VARIABLES

**Individuals** are the objects (or units) described by a set of data. Individuals may be people, but they may also be animals, plants, or things.

A **variable** is any one characteristic of an individual. A variable can take different values for different individuals.

A botanist's plant database, for example, includes data about various aspects of the plants examined. The plants are the individuals described by the database. For each individual (each plant), the data contain the values of variables such as stem length, number of flower petals, and flower color. An example of what such a database might look like is shown in [Figure 1.1](#), with each row dedicated to one individual and each column to a variable. **Spreadsheet** programs with rows and columns readily available are commonly used to store data and do simple calculations.



	A	B	C	D
1	Specimen ID	Stem length (cm)	Number of flower petals	Flower color
2	1	3.1	28	yellow
3	2	2.6	5	white
4	3	8.2	8	red
5	4	5.9	12	pink
6	5	14.7	5	yellow

**Figure 1.1**

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

**FIGURE 1.1** Example of a botanical database displayed in a spreadsheet. Each column includes data about a different variable. Each row represents data for one plant specimen.

Note that *sometimes the individuals represented in a data set are actually groups*, such as a set of countries, states, or counties, or even a set of ant colonies. For instance, every year, the U.S. Centers for Disease Control and Prevention (CDC) reports the percent of obese individuals in each state. In this data set, the 50 states are the individuals studied and the variable recorded for each individual is the percent of the population in that state who are obese.



## IDENTIFYING CATEGORICAL AND QUANTITATIVE VARIABLES

Some variables, like a plant's flower color, simply place individuals into categories. Others, like stem length and number of flower petals, take numerical values with which we can do arithmetic. It makes sense to give an average stem length for plants of a given environment, but it does not make sense to give an "average" flower color. We can summarize our findings on plant color by obtaining the counts of yellow, pink, and white flowers in the database, or by obtaining the proportion that each color type represents. *It is important, though, not to confuse these numerical summaries (average, count, or proportion) with the variables themselves.*



### CATEGORICAL AND QUANTITATIVE VARIABLES

A **categorical variable** places an individual into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense. The values of a quantitative variable are usually recorded in a **unit of measurement** such as seconds or kilograms.

Further distinctions are sometimes made in variables beyond simply categorical or quantitative. Some quantitative variables, like stem length, are **continuous** variables that can take any real numerical value over an interval. **Discrete** variables, in contrast, are quantitative variables that can take only a limited, finite number of values, like the number of petals in a flower.<sup>1</sup> Categorical variables can also be broken down into nominal and ordinal variables. **Nominal** variables are purely qualitative and unordered, like flower color, whereas **ordinal** data can be ranked, like star ratings or the Likert scales commonly used in psychology (for example, "Do you strongly disagree, disagree, agree, or strongly agree with the following statement?"). Although ordinal data can be ranked, they are not true quantitative variables, because the intervals between consecutive ranks are often not identical.

Be sure to carefully identify the individuals in the study first, so that you can determine what was recorded for each individual. For example, when the CDC reports the percent of obese individuals in each state, the individuals studied are the 50 states—not people. While people are either obese or not obese, each *state* provides a meaningful numerical value (the percent of its population who are obese). Therefore, the variable here is "percent of the population who are obese," and it is a quantitative variable.

### EXAMPLE 1.1 Paw preference in tree shrews

Tree shrews, *Tupaia belangeri*, are small omnivorous mammals that are phylogenetically related to primates. A research team examined paw preference during grasping tasks among 36 tree shrews born and raised in captivity. Here are some of the data they reported:<sup>2</sup>

Subject	Sex	Age	PI	Bias
Abel	m	6	-1.00	L
Anna	f	5	-0.95	L
Aragorn	m	6	-1.00	L
Barbossa	m	1	-0.25	A
Bea	f	4	1.00	R
Beatrice	f	1	0.56	R
Berta	f	1	0.13	A
...				

PRINTED BY: Olivia Mah (ommah@usfca.edu). Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



Oscar Dominguez/age fotostock/  
SuperStock



UNIVERSITY  
ommah@usfca.edu

As [Example 1.2](#) illustrates, it is important not to jump to conclusions. While words like “number,” “count,” “proportion,” or “percent” may be used to summarize categorical data, seeing these words does not necessarily imply that the variable recorded is categorical.

The publication referenced in [Example 1.2](#) presented a lot more findings. This is not particularly unusual. *Some scientific publications combine the findings from a series of related but distinct studies.* In such cases, you need to identify each study before you can determine the specific individuals and variables addressed in the study.



## APPLY YOUR KNOWLEDGE

**1.1 Cereal content.** Here is a small part of an ESEEE data set (available on the companion website), “Nutrition and Breakfast Cereals,” that describes the nutritional content per serving of 77 brands of breakfast cereal:<sup>4</sup>

Brand name	Manufacturer	Cold/Hot	Calories (Cal)	Sugar (grams)	Fiber (grams)
All Bran	K	C	70	5	9
All Bran with Extra Fiber	K	C	50	0	14
Almond Delight	R	C	110	8	1
Apple Cinnamon Cheerios	G	C	110	10	2
Apple Jacks	K	C	110	14	1
:					

- What are the individuals in this data set?
- For each individual, what variables are given? Which of these variables are categorical and which are quantitative?

**1.2 Concussion and sleep.** A 2016 National Public Radio (NPR) report described a study comparing patients who had their first ever traumatic brain injury 18 months earlier with similar healthy individuals with no prior brain trauma.<sup>5</sup> The news report states that, 18 months after their traumatic brain injury, the patients “were still getting, on average, an hour more sleep each night than similar healthy people were getting. And despite the extra sleep, 67 percent showed signs of excessive daytime sleepiness. Only 19 percent of healthy people had that problem.”

- Identify the individuals in the study.
- Identify the variables recorded and whether they are categorical or quantitative.



mark@rocketclips.com/Depositphotos

## CATEGORICAL VARIABLES:

### Pie charts and bar graphs

Statistical tools and ideas help us examine data so we can describe their main features. This examination is called **exploratory data analysis**. Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two principles that help us organize our exploration of a set of data.

#### EXPLORING DATA

1. Begin by examining each variable by itself. Then move on to study the relationships among the variables.
2. Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will also follow these principles in organizing our learning. [Chapters 1](#) and [2](#) present methods for describing a single variable. We study relationships among several variables in [Chapters 3](#) to [5](#). In each case, we begin with graphical displays, then add numerical summaries for more complete description.

The proper choice of graph depends on the nature of the variable. To examine a single variable, we usually want to display its *distribution*.

#### DISTRIBUTION OF A VARIABLE

The **distribution** of a variable tells us what values it takes and how often it takes these values.

The values of a categorical variable are labels for the categories. The **distribution of a categorical variable** lists the categories and gives either the count or the percent of individuals that fall in each category.

### EXAMPLE 1.3 Infectious diseases

The California Health and Human Services offers an open data portal (at <https://chhs.data.ca.gov>) providing state data on a variety of topics. You can, for instance, access all cases of infectious diseases reported in California. The database shows that, in 2014, California had a total of 262,780 reported infectious cases. Here is how these cases break down by type of infectious disease, after lumping the rarer diseases into an “other” category:

Infectious disease	Number of cases	Percent of all cases
Chlamydia	174,557	66
Gonorrhea	44,974	17
Pertussis	11,219	4
Campylobacteriosis	7,919	3
Early syphilis	7,191	3
Salmonellosis	5,361	2
Other	11,559	4
Total	262,780	100

The data table provides both the number of reported cases for each type of infectious disease and the percent that each disease represents in the set. Counts are also sometimes referred to as **frequencies**, and percents as **relative frequencies**. This table lists only the most common infectious diseases and lumps the remaining cases in the “other” category. Note that in 2014 California experienced an unusually large

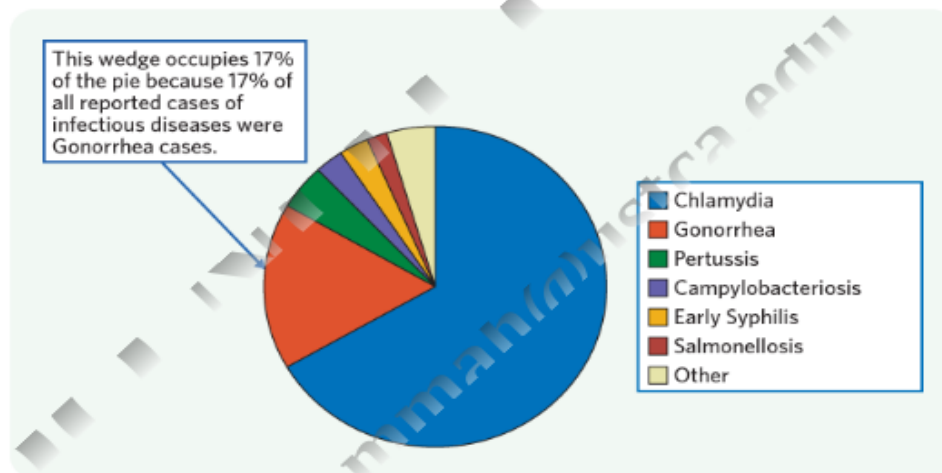


epidemic of pertussis, also known as whooping cough. Pertussis was the third most common infectious disease reported that year.

It's a good idea to check data for consistency. In this data set, the number of cases should add to 262,780, the total number of infectious disease cases reported in California in the year 2014. They do. The percents should add to 100% or, because each percent in the table is rounded to the nearest integer, very nearly 100%. Here the rounded percents add to 99 rather than 100. **Roundoff errors** don't point to mistakes in our work, just to the effect of rounding off results.



Columns of numbers take time to read and interpret. The **pie chart** in [Figure 1.2](#) shows the distribution of infectious cases more vividly. For example, the “gonorrhea” slice makes up 17% of the pie because 17% of all reported cases of infectious diseases in California in 2014 were gonorrhea cases. Pie charts are awkward to create by hand, but software will quickly do the job for you. *A pie chart can represent only one variable in one group at a time and must include all the categories that make up a whole.* Use a pie chart when you want to emphasize each category's relation to the whole. Here the cases all fall into either one of the six most common infectious diseases categories or the “other” category that lumps together all other infectious diseases. The graph shows more clearly than the table of raw numbers the overwhelming predominance of chlamydia and gonorrhea, two sexually transmitted diseases.



**Figure 1.2**

Baldi/Moore, *The Practice of Statistics in the Life Sciences, 4e*, © 2018 W. H. Freeman and Company

**FIGURE 1.2** You can use either a pie chart or a bar graph to display the distribution of a categorical variable. Here is a pie chart of the reported cases of infectious diseases in California for 2014.

We could also make a **bar graph** that represents each infectious disease through the height of a bar. Bar graphs are particularly adept at pointing out the order and the relative importance of the different categories. [Figure 1.3](#) shows two possible bar graphs of the data in [Example 1.3](#). The bar graph shown in [Figure 1.3\(a\)](#) is sorted by decreasing order of magnitude: The tallest bar appears first, followed by the second-tallest bar, and so on. It makes very clear that chlamydia and gonorrhea were by far the most common cases of infectious diseases reported in California in 2014. Because the data here are categorical, the bars in the graph may be arranged in any order we choose. Although the result would not be particularly interesting, the bars could be sorted alphabetically. Alternatively, the