Unit 3: Histograms

SUMMARY OF VIDEO

Many people are afraid of getting hit by lightning. And while getting hit by lightning is against the odds, it is not against all odds. Hundreds of people are struck by lightning every year in the U.S. What's more, fires started by lightning strikes cause hundreds of millions of dollars of property damage. Meteorologist Raul Lopez and his associates began collecting detailed data on lightning strikes back in the 1980s and soon were overwhelmed by the vast amount of data. In one year, they collected three-quarters of a million flashes in a small area of Colorado. They decided to focus on when lightning strikes occurred. The data on the times of the first lightning strike needed to be organized, summarized, and displayed graphically. One of the statistical tools that Raul Lopez turned to was the graphic display called a histogram. For example, data on the percent of first lightning flashes for each hour of the day is displayed in the histogram in Figure 3.1.

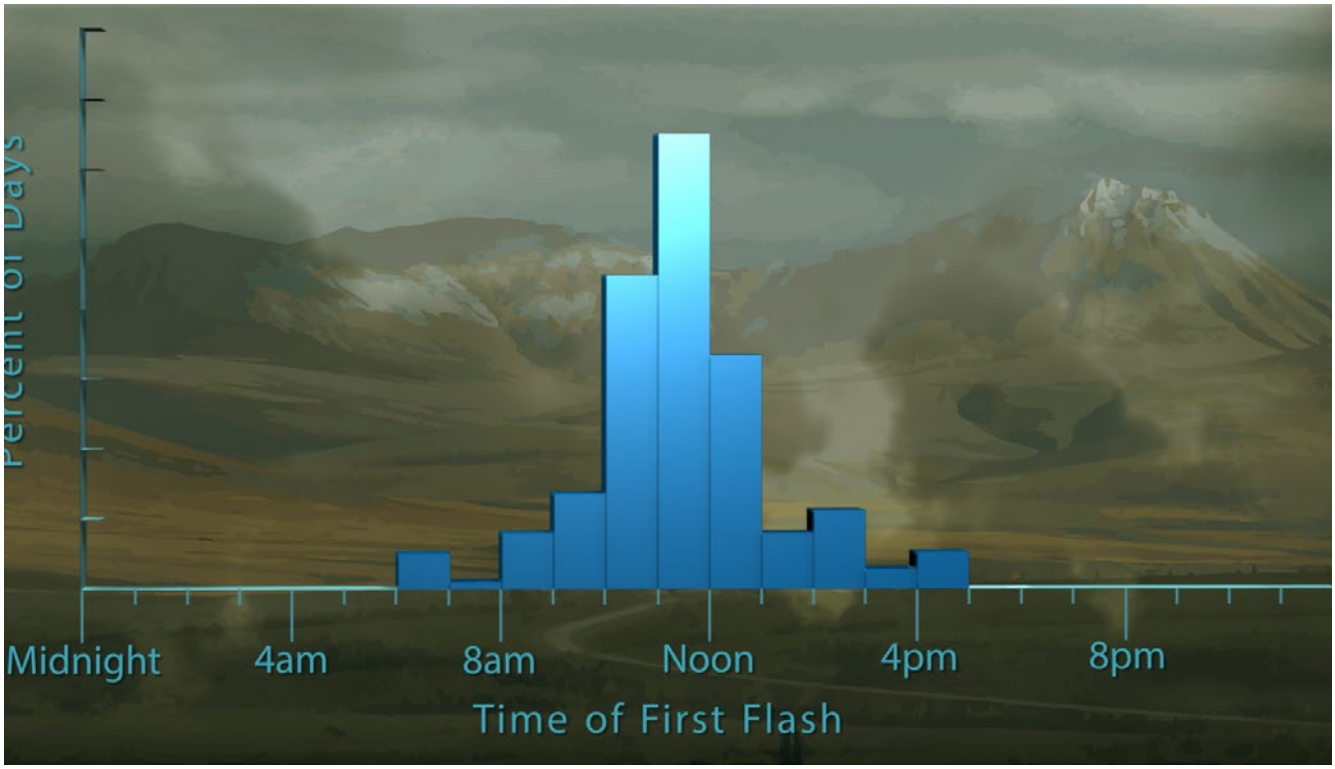

Figure 3.1. Histogram of the time of the first lightning strike.

Before the histogram could be constructed, each day was broken into hours (horizontal axis), the number of first flashes in each hour was counted, and then the counts were converted to percentages (vertical axis). So, in this histogram, each bar represents one hour, and its height is the percentage of days in which the first lightning flash fell in that hour. This histogram has two very striking features. First, it is roughly symmetric about the tallest bar, which represents the percentage of first flashes between 11 a.m. and noon. The second rather surprising feature is how tightly the time of first strike clusters around the center bar, with a range from 10 a.m. to 1 p.m. accounting for most of the days' first strikes. And there are no first strikes at night. This pattern helped explain how lightning storms form in this area. This region is mountainous and winds from the eastern plains carry warm moist air. When the wind hits the mountains it is forced upward where it meets and mixes with colder air higher in the atmosphere forming clouds. And this turns out to be a regular daily occurrence during the Colorado summer.

Lopez and his colleagues next looked at the time of day when the maximum number of lightning flashes occurred. (See Figure 3.2.) They found a similar pattern, with a peak showing that most flashes occur between 4 p.m. and 5 p.m.

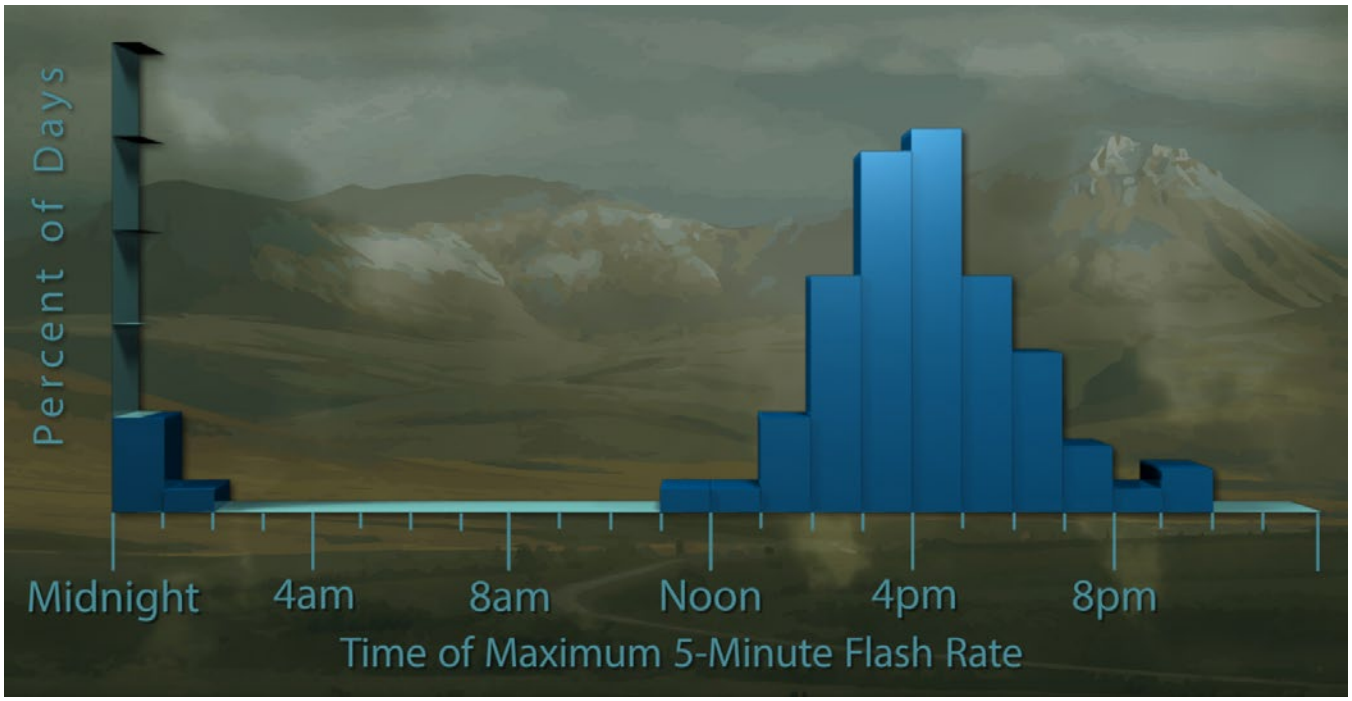

Figure 3.2. Histogram of the time of maximum flash rate.

But there is one big difference from the first flash histogram in Figure 3.1. On a few days the maximum was in the early hours of the morning. Data points like these, which stand out from the overall pattern of the distribution, are called outliers. Outliers are often the most intriguing features of a histogram. Outliers should always be investigated and, if possible, explained.

The explanation that Lopez and his colleagues came up with was that they occur on days when larger weather systems, specifically very strong winds from fast moving weather fronts, overpower the local effect.

Data collection on Colorado lightning has continued since the pioneering work of Raul Lopez and his colleagues. Figure 3.3 shows a histogram produced from more recent data showing the number of people injured or killed by lightning strikes in the last 30 years. It shows the same clustering pattern as Raul Lopez's histograms, but interestingly, the peak time for getting struck by lightning is around 2 p.m., about midway between the peaks of the first strike and maximum activity histograms.

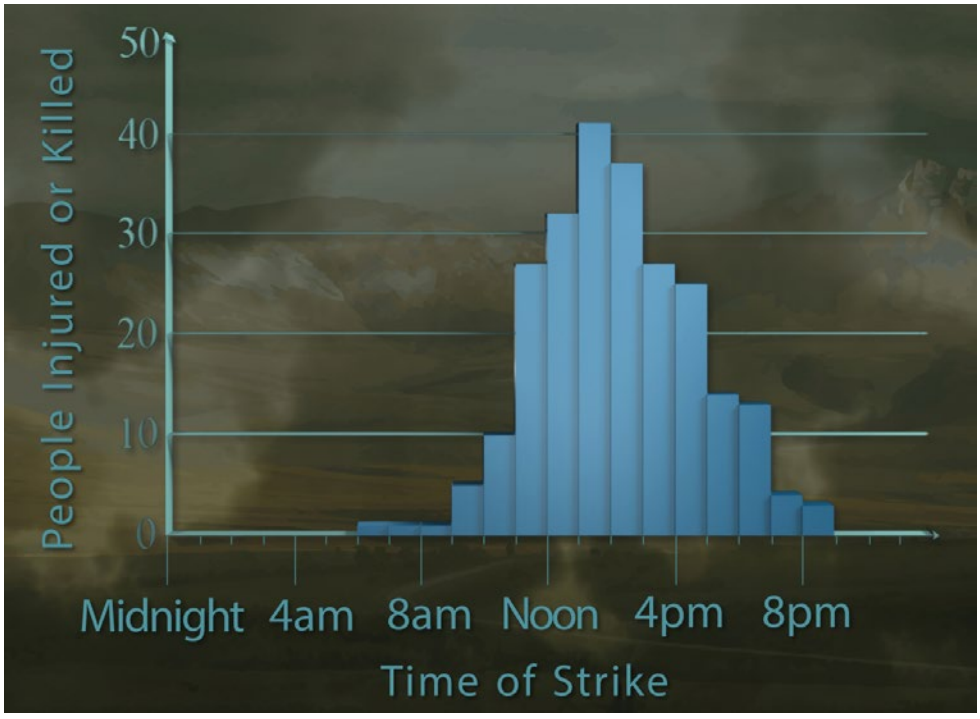

Figure 3.3. Histogram of time when people were struck by lightning.

When constructing histograms it is very important to choose the best class size – that is, the choice of the interval widths for the horizontal axis. Lopez chose one hour for his data, and it works well. But suppose we turn our attention to a different context, the weekday traffic density on a portion of the Massachusetts Turnpike. First, we look at a histogram with class intervals of three hours. (See Figure 3.4.)

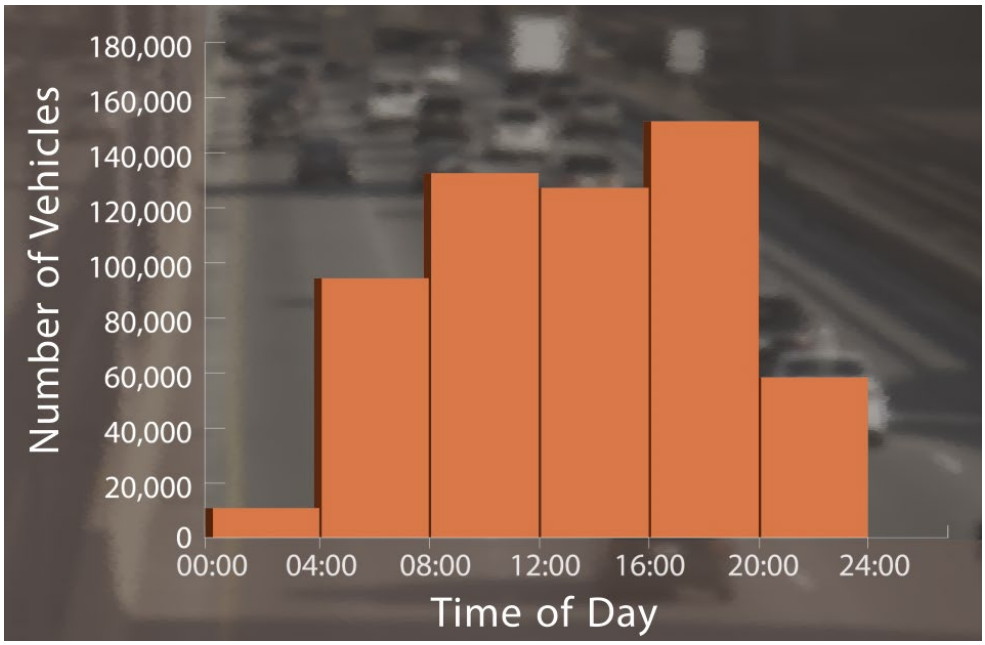

Figure 3.4. Histogram of traffic density in three-hour intervals.

The histogram in Figure 3.4 is not terribly informative. Next, we changed the interval width to one hour, which was better. However, using one-half hour widths as shown in Figure 3.5 is even better. Now, the increased traffic density during morning rush hour and evening rush hour is clearly visible in the pattern of two peaks.

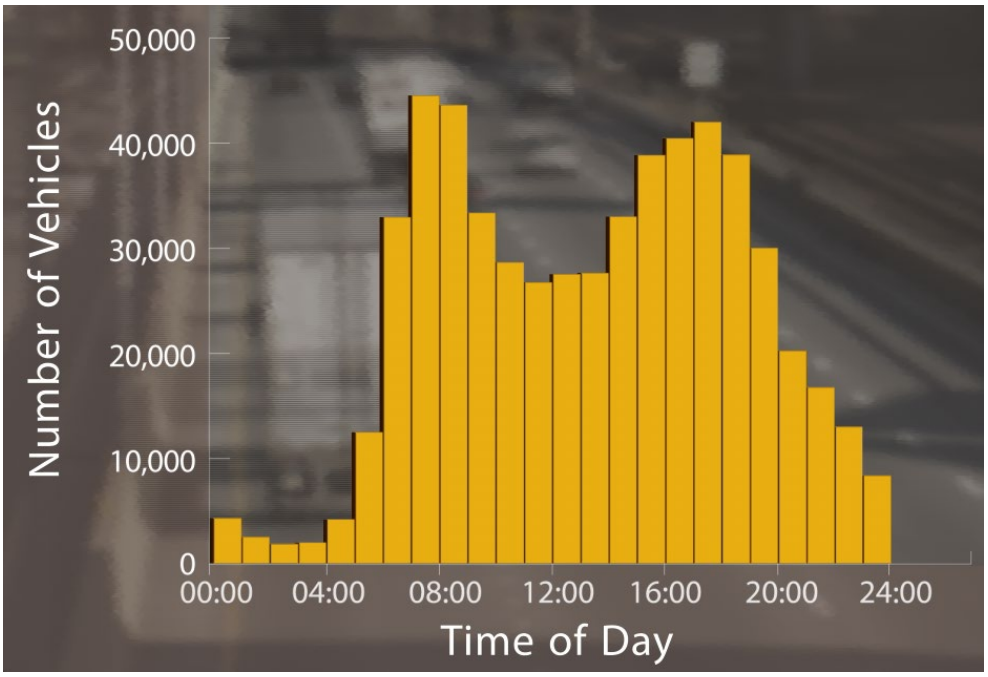

Figure 3.5. Histogram of traffic density in half-hour intervals.

But what if we went even finer-grained and used 5-minute intervals? Take a look at Figure 3.6. Now the peaks begin disappearing again back into the numbers and the histogram becomes less informative.

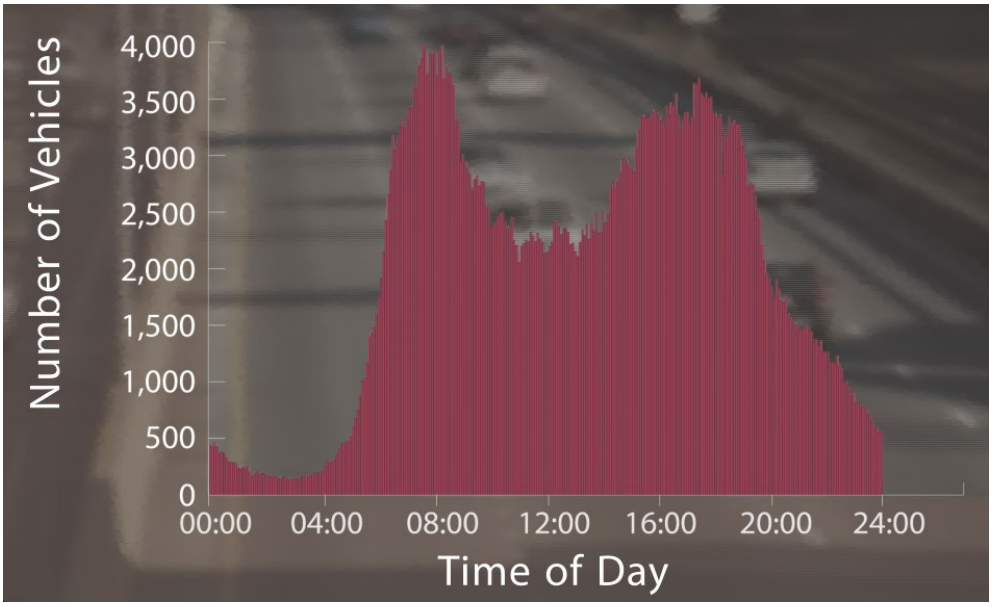

Figure 3.6. Histogram of traffic density in 5-minute intervals.

So, we have seen how histograms can literally show at a glance the essence of a whole lot of numbers. Here is one last example. Figure 3.7 shows a histogram of the weekly wages of workers in the U.S. in the year 1992.

Figure 3.7. Histogram of weekly wages (1992).

Notice how strikingly it is skewed, with most people earning around $450 per week. As you go out to what is called the tail of the distribution (to the right), the salaries get bigger, but the

percent of people earning those salaries gets smaller. Statisticians say a distribution like this is skewed to the right, because the right side of the histogram extends much further out than the left side. Now look at the histogram in Figure 3.8 of the same variable, weekly wages, but for the year 2011.

Figure 3.8. Histogram of weekly wages (2011).

Now, the skew has become much more pronounced, and the tail has grown much longer. Suddenly our little discourse on histograms could become highly political!

STUDENT LEARNING OBJECTIVES

A. Understand that the distribution of a variable consists of what values the variable takes and how often. (This is a repeat of an objective from Unit 2, Stemplots.)

B. Be able to construct a histogram to display the distribution of a variable for moderate amounts of data (say, data sets with fewer than 200 observations).

C. Understand that class intervals should be of equal width; choose appropriate class widths to effectively reveal informative patterns in the data.

D. Understand that the vertical axis of the histogram may be scaled for frequency, proportion, or percentage. The choice of vertical scaling for any data set does not affect the important features revealed by a histogram.

E. Be able to describe a graphical display of data by first describing the overall pattern and then deviations from that pattern. Describe the shape of the overall pattern and identify any gaps in data and potential outliers.

F. Recognize rough symmetry and clear skewness in the overall pattern of a distribution.

CONTENT OVERVIEW

Rows and rows of data provide little information. For example, below are thickness measurements, in millimeters, from a sample of 25 polished wafers used in the manufacture of microchips. Notice that it is difficult to extract much information from staring at these numbers. The numbers need to be organized, summarized, and displayed graphically in order to unlock the information they contain.

0.402	0.496	0.533	0.387	0.384
0.528	0.411	0.367	0.462	0.499
0.539	0.546	0.425	0.457	0.586
0.558	0.588	0.425	0.437	0.479
0.427	0.485	0.443	0.441	0.658

A **frequency distribution** is one method of organizing and summarizing data in a table. The basic idea behind a frequency distribution is to set up categories (class intervals), classify data values into the categories, and then determine the frequency with which data values are placed into each category. The steps below outline the process of making a frequency distribution table.

Creating a frequency distribution table

Step 1: Identify an interval that is wide enough to contain all the data.

Step 2: Subdivide the interval identified in Step 1 into class intervals of equal width. The class intervals will serve as the categories.

Step 3: Set up a table with three columns for the following: class interval, tally, and frequency. (The tally column can be removed in the final table.)

Step 4: To complete the table, determine the frequency with which data values fall into each class interval.

Convention: Any data value that falls on a class interval boundary is placed in the class interval to the right. If the data value is a maximum, it is generally put in the interval that contains the maximum at its right endpoint.

Now, we apply Steps 1 – 4 to make a frequency distribution table for the thickness measurements.

Step 1: In this case the smallest data value is 0.367 mm and the largest is 0.698 mm. We choose the interval from 0.3 mm to 0.7 mm, which contains all the thickness measurements.

Step 2: The total width of the interval from 0.3 to 0.7 is 0.4. Dividing this interval into eight class intervals works out nicely – each class interval will have width 0.05.

Step 3: We have set up Table 3.1 to have three columns, which we have labeled Thickness, Tally, and Frequency. We have entered the endpoints of the eight class intervals into the Thickness column.

Thickness (mm)	Tally	Frequency
0.30 – 0.35		
0.35 – 0.40		
0.40 – 0.45		
0.45 – 0.50		
0.50 – 0.55		
0.55 – 0.60		
0.60 – 0.65		
0.65 – 0.70		

Table 3.1: Setting up a frequency distribution table.

Step 4: The easiest way to determine the frequencies is to draw a tally line for each data value that falls into a particular class interval. When drawing tally lines, keep the following in mind:

- As you draw tally lines, instead of drawing a fifth tally line, cross out the previous four.
- If a data value falls on the boundary of a class interval, record it in the interval with the larger values.

Once a tally line has been drawn for each data value, count the number of tally lines corresponding to each class interval and record that number in the frequency column as shown in Table 3.2.

Thickness (mm)	Tally	Frequency
0.30 – 0.35		0
0.35 – 0.40	|||	3
0.40 – 0.45	𝍸 |||	8
0.45 – 0.50	𝍸 |	6
0.50 – 0.55	||||	4
0.55 – 0.60	|||	3
0.60 – 0.65		0
0.65 – 0.70	|	1

Table 3.2: A completed frequency distribution table.

The frequency distribution in Table 3.2 reveals more information about the data than a quick look at the 25 numbers. For example, from the frequency distribution, we learn that more measurements fall in the interval 0.40 – 0.45 than in any of the other class intervals. Also, we learn there is a gap in the data – no data values fall between 0.60 and 0.65.

Although a frequency distribution table is a useful tool for extracting information from data, a histogram can often convey the same information more effectively. Next, we outline how to construct a histogram from a frequency distribution.

Creating a histogram from a frequency distribution

Step 1: Draw a set of axes. On the horizontal axis, mark the boundaries of the class intervals. On the vertical axis, set up a scale appropriate for the frequencies. (Later this scale can be changed to proportion or percent.)

Step 2: Label the horizontal axis with the name of the variable being measured and the units.

Step 3: Over each interval, draw a rectangle with the interval as its base. The height of the rectangle should match the frequency of data contained in that interval.

Next, we apply Steps 1 – 3 for creating a histogram to the frequency distribution in Table 3.2. Figure 3.9 shows the results.

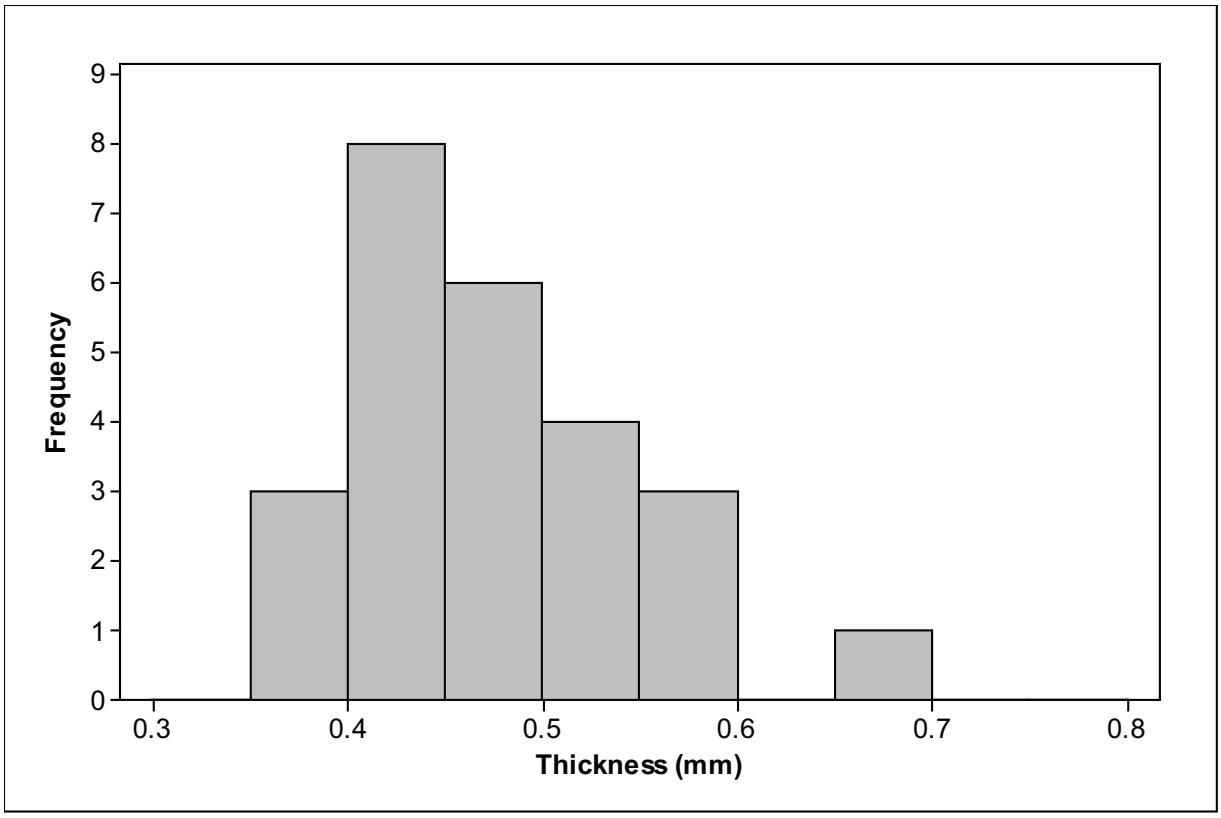

Figure 3.9: Histogram representing frequency distribution in Table 3.2.

Particularly if you are comparing histograms from samples with a different number of data values, it is useful to replace the frequency scale on the vertical axis with the proportion or percent.

Calculating Proportions and Percents

- To calculate a proportion, divide the frequency by the sample size.
- To convert a proportion into a percent, multiply the proportion by 100%.

In describing a histogram, we first look for the overall pattern of the distribution. In sizing up the overall pattern, look for the following:

- center and spread;
- one peak or several (unimodal or multimodal);
- a regular shape, such as symmetric or skewed.

In the case of the histogram in Figure 3.9, the overall pattern is single-peaked (or unimodal) and skewed to the right. Next, we look for any striking deviations from that pattern. An important kind of deviation from an overall pattern is an outlier, an individual observation

that lies clearly outside the overall pattern. Once identified, outliers should be investigated. Sometimes they are errors in the data and sometimes they have interesting stories related to the data. For Figure 3.9, there is a gap between 0.6 and 0.65 and there is one data value between 0.65 and 0.70, which might be an outlier.

KEY TERMS

A **frequency distribution** provides a means of organizing and summarizing data by classifying data values into class intervals and recording the number of data that fall into each class interval.

A **histogram** is a graphical representation of a frequency distribution. Bars are drawn over each class interval on a number line. The areas of the bars are proportional to the frequencies with which data fall into the class intervals.

The shape of a unimodal distribution of a quantitative variable may be **symmetric** (right side close to a mirror image of left side) or skewed to the right or left. A distribution is **skewed to the right** if the right tail of the distribution is longer than the left and is **skewed to the left** if the left tail of the distribution is longer than the right.

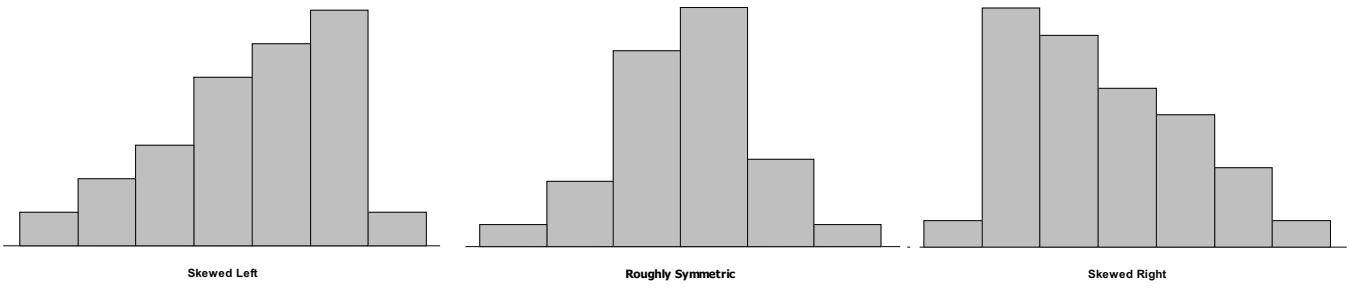

Figure 3.10. Shapes of histograms.