Chapter 11

Displaying Distributions with Graphs

Lecture Slides

Case Study: Displaying Distributions with Graphs 1

Nutritionists tell us that a healthy diet should include 20 to 35 grams of fiber daily.

Cereal manufacturers advertise their products as “high-fiber.”

The food label on the side of a box of cereal (mandated by the Food and Drug Administration) provides information that allows the consumer to choose a healthy breakfast cereal.

You will find lots of different cereals displayed at the grocery store.

Case Study: Displaying Distributions with Graphs 2

You could examine all the boxes to see how much fiber each contains, but how do you make sense of all the numbers?

Is your favorite cereal, Wheaties, with 3 grams of dietary fiber, among those with the highest fiber content? How will you choose?

A histogram or stemplot could help. By the end of this chapter, you will know how to make a histogram and stemplot and know what to look for when you study one of these graphs.

Histograms 1

Categorical variables record group membership, such as the marital status of a man or the race of a college student.

What about quantitative variables such as the SAT scores of students admitted to a college or the income of families?

These variables take so many values that a graph of the distribution is clearer if nearby values are grouped together. The commonest graph of the distribution of a quantitative variable is a **histogram**.

Example: How to make a histogram 1

Table 11.1 presents the percent of residents aged 65 years and over in each of the 50 states.

Table 11.1 Percentage of residents aged 65 and older in states, 2010

State	Percent	State	Percent	State	Percent
Alabama	13.8	Louisiana	12.3	Ohio	14.1
Alaska	7.7	Maine	15.9	Oklahoma	13.5
Arizona	13.8	Maryland	12.3	Oregon	13.9
Arkansas	14.4	Massachusetts	13.8	Pennsylvania	15.4
California	11.4	Michigan	13.8	Rhode Island	14.4
Colorado	10.9	Minnesota	12.9	South Carolina	13.7
Connecticut	14.2	Mississippi	12.8	South Dakota	14.3
Delaware	14.4	Missouri	14.0	Tennessee	13.4
Florida	17.3	Montana	14.8	Texas	10.3
Georgia	10.7	Nebraska	13.5	Utah	9.0
Hawaii	14.3	Nevada	12.0	Vermont	14.6
Idaho	12.4	New Hampshire	13.5	Virginia	12.2
Illinois	12.5	New Jersey	13.5	Washington	12.3
Indiana	13.0	New Mexico	13.2	West Virginia	16.0
Iowa	14.9	New York	13.5	Wisconsin	13.8
Kansas	13.2	North Carolina	12.9	Wyoming	12.4
Kentucky	13.3	North Dakota	14.5		

Data from *Age and Sex Composition: 2010 Census Briefs;* available online at https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf.

Example:
How to make a histogram 2

To create a histogram of the data in Table 11.1, do the following:

Step 1. Divide the range of the data into classes of equal width. The data in Table 11.1 range from 7.7 to 17.3, so we choose as our classes

$7.0 \le$ percentage over 65 < 8.0
$8.0 \le$ percentage over 65 < 9.0
$17.0 \le$ percentage over 65 < 18.0

Be sure to specify the classes precisely so that each individual falls into exactly one class.

Example:
How to make a histogram 3

Step 2. Count the number of individuals in each class.
Here are the counts:

Class	Count
7.0 to 7.9	1
8.0 to 8.9	0
9.0 to 9.9	1
10.0 to 10.9	3

Class	Count
11.0 to 11.9	1
12.0 to 12.9	11
13.0 to 13.9	17
14.0 to 14.9	12

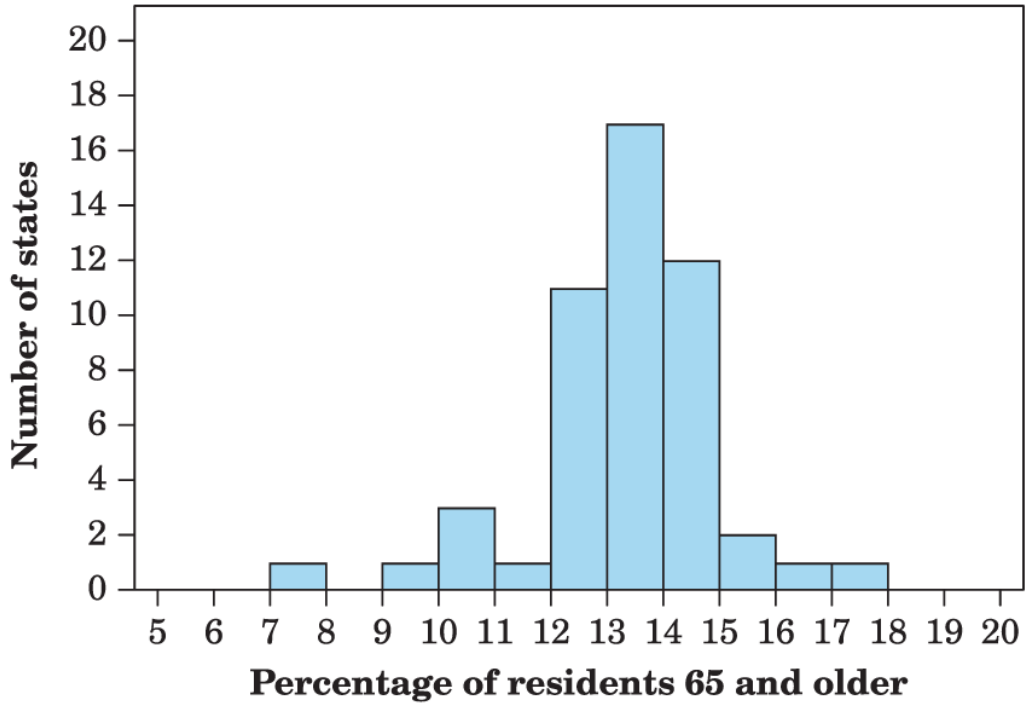
Class	Count
15.0 to 15.9	2
16.0 to 16.9	1
17.0 to 17.9	1

Example:
How to make a histogram 4

Step 3. Draw the histogram. Mark on the horizontal axis the scale for the variable whose distribution you are displaying. That's "Percentage of residents aged 65 and over" in this example. The scale runs from 5 to 20 because that range spans the classes we chose. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty so that its bar has height zero.

Example:
How to make a histogram 5

Figure 11.1 is our histogram.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Histograms 2

The classes for a histogram should have equal widths. Choose the number of classes wisely.

Some people recommend between 10 and 20 classes but suggest using fewer when the size of the data set is small.

Too few classes will give a “skyscraper” histogram, with all values in a few classes with tall bars.

Too many classes will produce a “pancake” graph, with most classes having one or no observations.

Histograms 3

Use your judgment in choosing classes to display the shape of a distribution.

Statistics software will choose the classes for you and may use slightly different rules than those we have discussed.

The computer's choice is usually a good one, but you can change it if you want. When using statistical software, it is good practice to check which rules are used to determine the classes.

Interpreting Histograms 1

Making a statistical graph is not an end in itself. The purpose of the graph is to help us understand the data.

After you (or your computer) make a graph, always ask, "What do I see?" Here is a general strategy for looking at graphs.

In any graph of data, look for an **overall pattern** and also for **striking deviations** from that pattern.

Interpreting Histograms 2

In the case of the histogram of Figure 11.1, it is easiest to begin with deviations from the overall pattern of the histogram.

Two states stand out as separated from the main body of the histogram. You can find them in the table once the histogram has called attention to them.

Alaska has 7.7% and Florida 17.3% of its residents over age 65. These states are clear **outliers**.

Interpreting Histograms 3

An **outlier** in any graph of data is an individual observation that falls outside the overall pattern of the graph.

Whether an observation is an outlier is to some extent a matter of judgment, although statisticians have developed some objective criteria for identifying possible outliers.

Look for an explanation for outliers, either an error or the special nature of some observations.

Interpreting Histograms 4

To see the overall pattern of a histogram, ignore any outliers.

Here is a simple way to organize your thinking.

To describe the **overall pattern** of a distribution:

- Describe the **center** and the **variability**.
- Describe the **shape** of the histogram in a few words

Interpreting Histograms 5

We will learn how to describe center and variability numerically in Chapter 12.

For now, describe the center of a distribution by its midpoint, the value at roughly the middle of all the values in the distribution.

Describe the variability of a distribution by giving the smallest and largest values, ignoring any outliers.

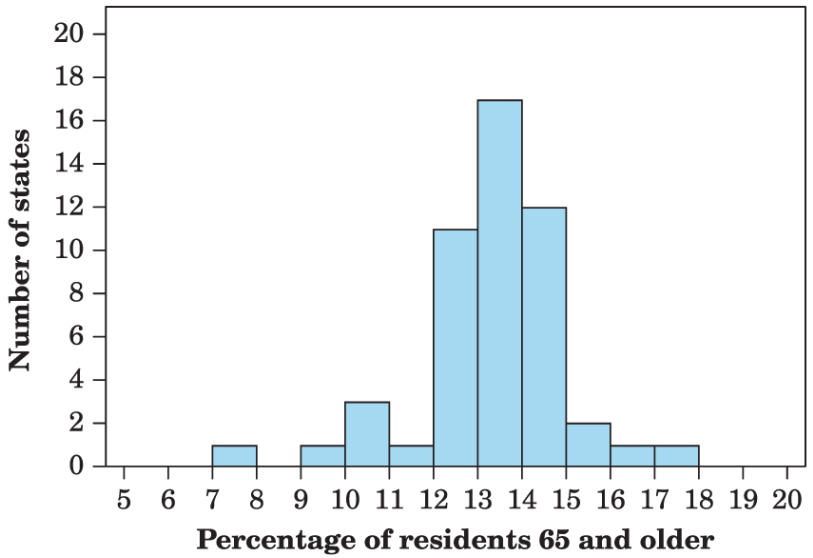
Example: Describing distributions

See Figure 11.1 to the right.

Shape: The distribution has a single peak. It is roughly symmetrical—that is, the pattern is similar on both sides of the peak.

Center: The midpoint of the distribution is close to the single peak, at about 13%.

Variability: The variability is about 9% to 18% if we ignore the outliers



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Interpreting Histograms 6

Symmetrical and skewed distributions

A distribution is **symmetrical** if the right and left sides of the histogram are approximately mirror images of each other.

A distribution is **skewed to the right** if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side.

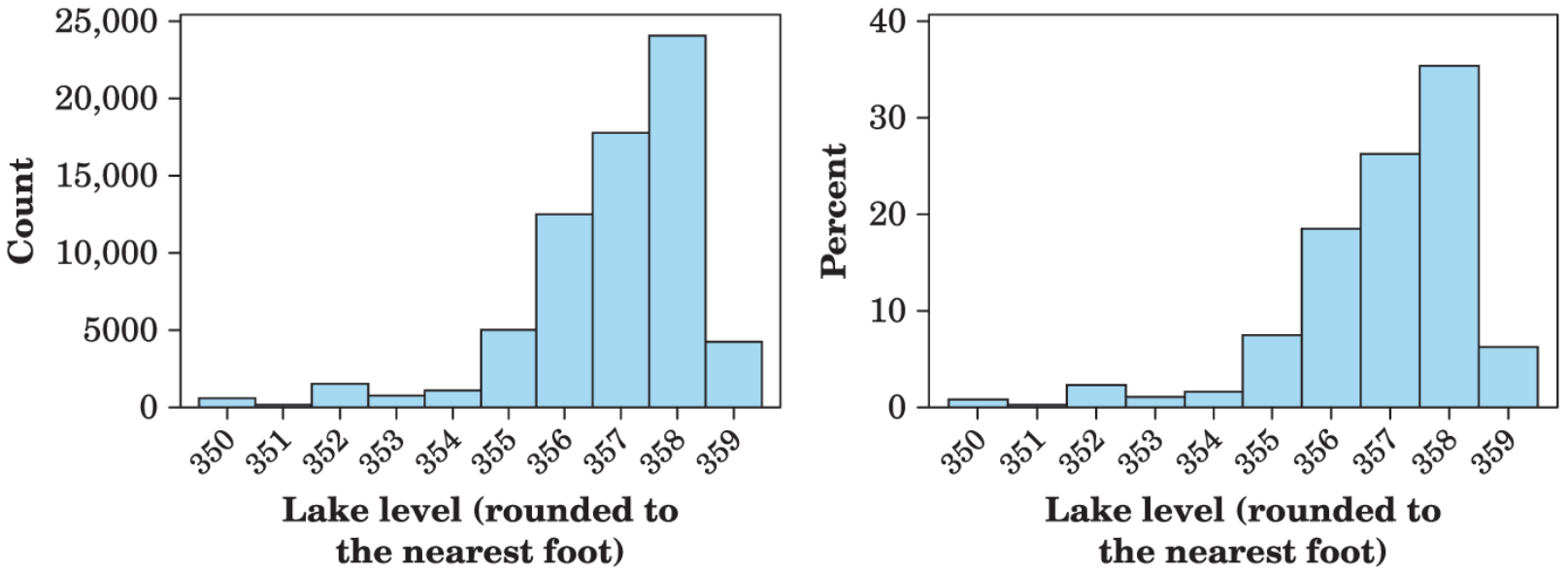
It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

Example: Lake elevation levels 1

Lake Murray is a man-made reservoir located in South Carolina. It is used mainly for recreation, such as boating, fishing, and water sports. It is also used to provide back-up hydroelectric power for South Carolina Electric and Gas. The lake levels fluctuate with the highest levels in summer (for safe boating and good fishing) and the lowest levels in winter (for water quality). Water can be released at the dam in the case of heavy rains or to let water out to maintain winter levels. The US Geological Survey (USGS) monitors water levels in Lake Murray.

Example: Lake elevation levels 2

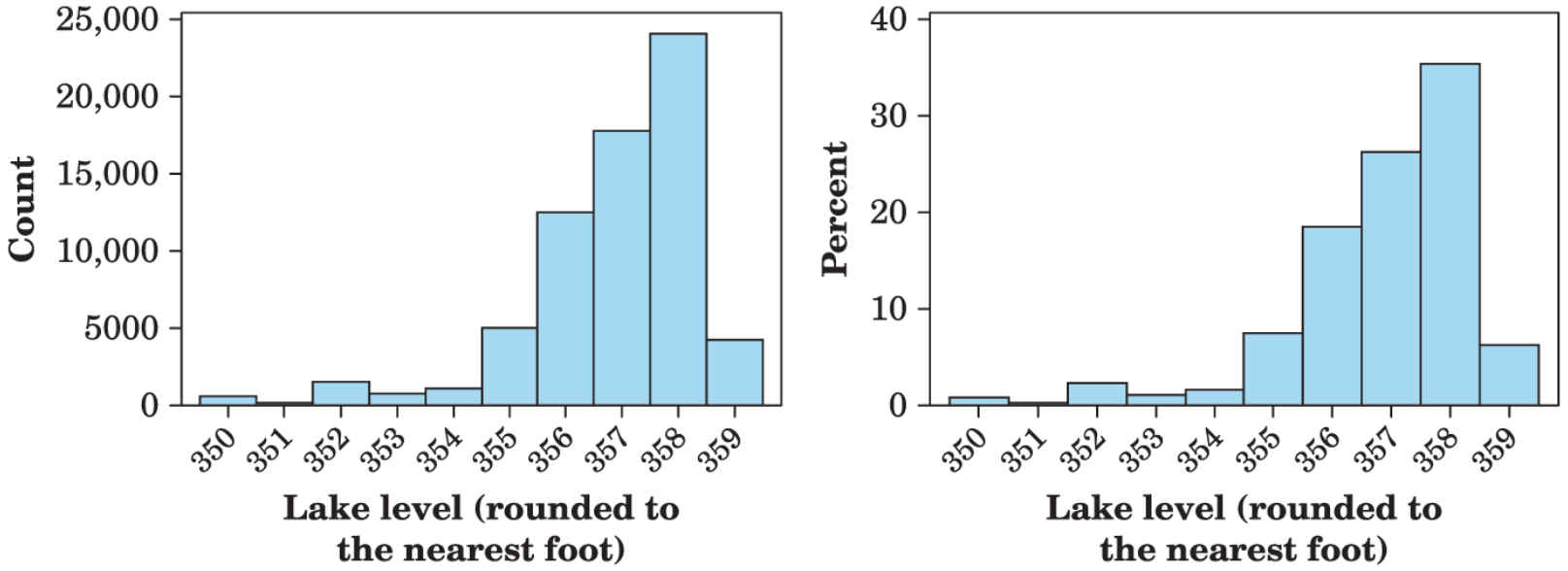
The histograms below were created using 67,810 hourly elevation levels for Lake Murray from November 1, 2007, through August 11, 2015. The two histograms of lake levels were made from the same data set, and the histograms look identical in shape.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Lake elevation levels 3

The shape of the distribution of lake levels is skewed left since the left side of the histogram is longer.

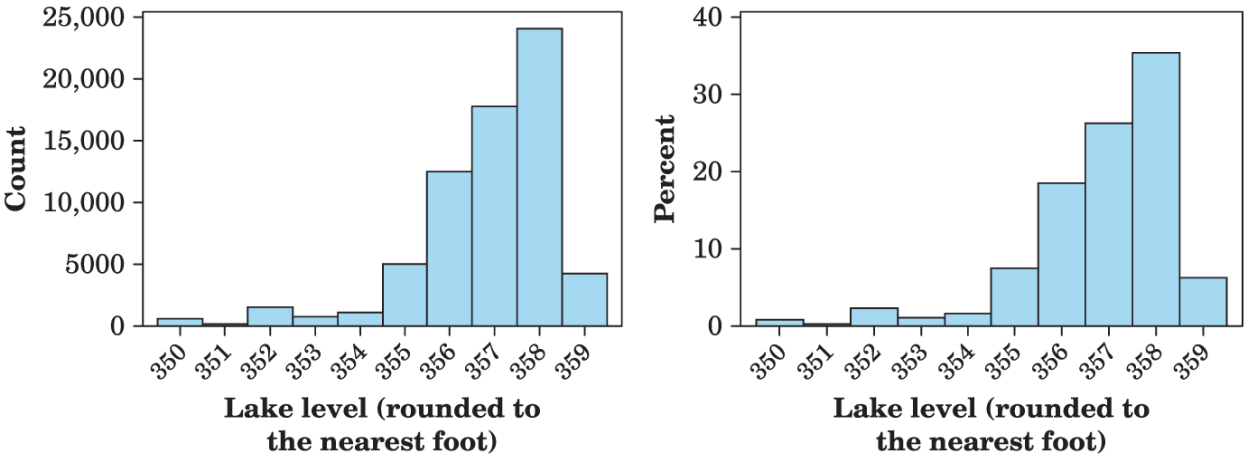


Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Lake elevation levels 4

Let's examine the difference in the two histograms.

The histogram on the left puts the count of observations on the *y* axis (this is called a frequency histogram), while the histogram on the right uses the percent of times the lake reaches a certain level (this is called a relative frequency histogram).



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Lake elevation levels 5

The frequency histogram tells us the lake reached an elevation of 358 feet approximately 24,000 times (24,041 to be exact!).

If a fisherman considering a move to Lake Murray cares about how often the lake reaches a certain level, it is more illustrative to use the relative frequency histogram on the right, which reports the percent of times the lake reached 358 feet. The height of the bar for 358 feet is 35, so the fisherman would know the lake is at the 358-foot elevation roughly 35% of the time.

Interpreting Histograms 7

It is not uncommon in the current world to have very large data sets.

Google uses big data to rank web pages and provide the best search results.

Banks use big data to analyze spending patterns and learn when to flag your debit or credit card for fraudulent use.

Large firms use big data to analyze market patterns and adapt marketing strategies accordingly.

Interpreting Histograms 8

Our data set of size 67,810 is actually small in the realm of “big data” but is still big enough to see that it is almost always better to use a relative frequency histogram when sample sizes grow large.

A relative frequency histogram is also a better choice if one wants to make comparisons between two distributions.

Stemplots 1

Histograms are not the only graphical display of distributions.

For small data sets, a **stemplot** (sometimes called a **stem-and-leaf plot**) is quicker to make and presents more detailed information

Stemplots 2

To make a stemplot:

1. Separate each observation into a stem consisting of all but the final (rightmost) digit and a leaf, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit. Do not include commas or decimal points with your leaves.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.

3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Example: Stemplot of the “65 and over” data 1

For the “65 and over” percentages in Table 11.1, the whole-number part of the observation is the stem, and the final digit (tenths) is the leaf. The Alabama entry, 13.8, has stem 13 and leaf 8.

Stems can have as many digits as needed, but each leaf must consist of only a single digit.

Example: Stemplot of the “65 and over” data 2

Stemplot of the “65 and over” data.

stem	leaf	stem	leaf	stem	leaf
7		7	7	7	7
8		8		8	
9		9	0	9	0
10		10	973	10	379
11		11	4	11	4
12		12	45339809234	12	02333445899
13		13	88023885552559748	13	02234555557888889
14		14	424390851436	14	012334445689
15		15	94	15	49
16		16	0	16	0
17		17	3	17	3

Key: A stem of 7 and a leaf of 7 means 7.7%

Example: Stemplot of the “65 and over” data 3

The chief advantage of a stemplot is that it displays the actual values of the observations.

Stemplots are faster to draw than histograms.

A stemplot requires that we use the first digit or digits as stems. This amounts to an automatic choice of classes and can give a poor picture of the distribution.

Stemplots do not work well with large data sets, because the stems then have too many leaves.

Statistics in Summary 1

- The **distribution** of a variable tells us what values the variable takes and how often it takes each value.

- To display the distribution of a quantitative variable, use a **histogram** or a **stemplot**. We usually favor stemplots when we have a small number of observations and histograms for larger data sets. Make sure to choose the appropriate number of classes so that the distribution shape is displayed accurately. For really large data sets, use a histogram of percents (relative frequency histogram).

Statistics in Summary 2

- When you look at a graph, look for an **overall pattern** and for **deviations** from that pattern, such as **outliers**.

- We can characterize the overall pattern of a histogram or stemplot by describing its **shape**, **center**, and **variability**. Some distributions have simple shapes such as **symmetrical**, **skewed left**, or **skewed right**, but others are too irregular to describe by a simple shape.