

CHAPTER 1

Exploring Data

1.2

Displaying Quantitative Data with Graphs

Dot plots

One of the simplest graphs to construct and interpret is a **dot plot**. Each data value is shown as a dot above its location on a number line.

How to make a dotplot:

Draw a horizontal axis (a number line) and label it with the variable name.

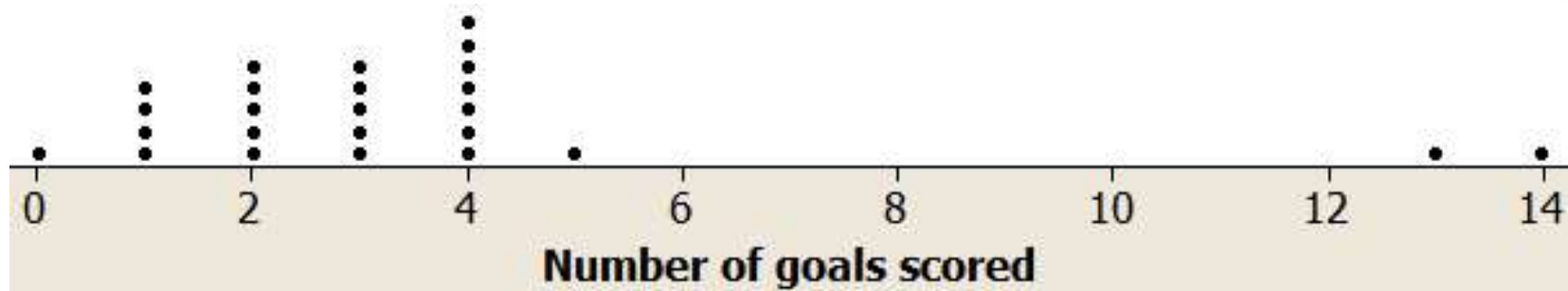
Scale the axis from the minimum to the maximum value.

Mark a dot above the location on the horizontal axis corresponding to each data value.

Example 1: How good was the 2012 U.S. women's soccer team? With players like Abby Wambach, Megan Rapinoe, and Hope Solo, the team put on an impressive showing en route to winning the gold medal at the 2012 Olympics in London. Here are data on the number of goals scored by the team in the 12 months prior to the 2012 Olympics.

1 3 1 14 13 4 3 4 2 5 2 0 4

1 3 4 3 4 2 4 3 1 2 4 2



Examining the Distribution of a Quantitative Variable

The purpose of a graph is to help us understand the data. After you make a graph, always ask, “What do I see?”

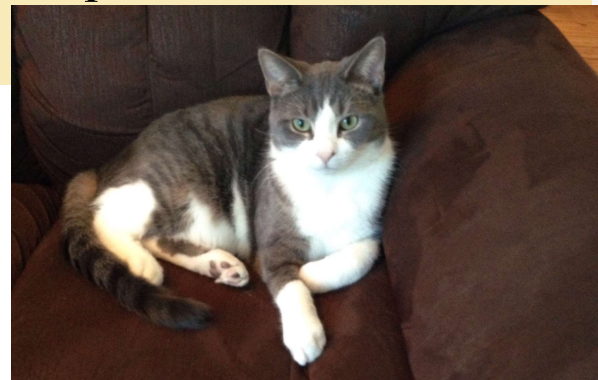
How to Examine the Distribution of a Quantitative Variable

1. In any graph, look for the **overall pattern** and for striking **departures** from that pattern.
2. Describe the overall pattern of a distribution by its:

- ***Shape**
- ***Center**
- ***Spread**

Don't forget your SOCS!

3. Note individual values that fall outside the overall pattern. These departures are called **outliers**.

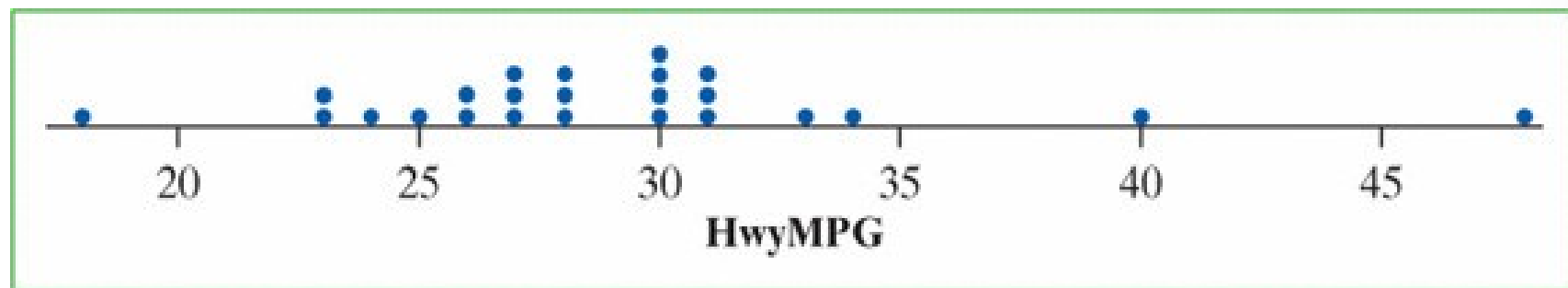


Example 2: The Environmental Protection Agency (EPA) is in charge of determining and reporting fuel economy ratings for cars (think of those large window stickers on a new car). For years, consumers complained that their actual gas mileages were noticeably lower than the values reported by the EPA. It seems that the EPA's tests—all of which are done on computerized devices to ensure consistency—did not consider things like outdoor temperature, use of the air conditioner, or realistic acceleration and braking by drivers. In 2008 the EPA changed the method for measuring a vehicle's fuel economy to try to give more accurate estimates.

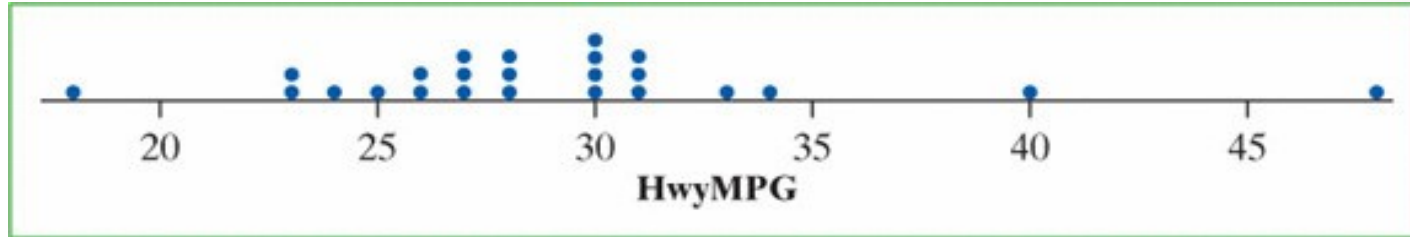
The table on the next slide displays the EPA estimates of highway gas mileage in miles per gallon (mpg) for a sample of 24 model year 2012 midsize cars.

Model	mpg	Model	mpg	Model	mpg
Acura RL	24	Dodge Avenger	30	Mercedes-Benz E350	30
Audi A8	28	Ford Fusion	25	Mitsubishi Galant	30
Bentley Mulsanne	18	Hyundai Elantra	40	Nissan Maxima	26
BMW 550i	23	Jaguar XF	23	Saab 9-5 Sedan	28
Buick Lacrosse	27	Kia Optima	34	Subaru Legacy	31
Cadillac CTS	27	Lexus ES 350	28	Toyota Prius	48
Chevrolet Malibu	33	Lincoln MKZ	27	Volkswagen Passat	31
Chrysler 200	30	Mazda 6	31	Volvo S80	26

The figure below shows a dotplot of the data:



Describe the shape, center, and spread of the distribution. Are there any outliers?



Shape: The dotplot has a peak at 30 mpg and a main cluster of values from 23 to 34 mpg. There are large gaps between 18 and 23, 34 and 40, 40 and 48 mpg.

Center: The midpoint of the 24 values shown in the graph is 28. So a typical model year 2012 midsize car in the sample got about 28 miles per gallon on the highway.

Spread: The data vary from 18 mpg to 48 mpg. The 2012 Nissan Leaf, an electric car, got an EPA estimated 92 miles per gallon on the highway. With the U.S. government's plan to raise the fuel economy standard to an average of 54.5 mpg by 2025, even more alternative-fuel vehicles like the Leaf will have to be developed.

Outliers: We see two midsize cars with unusually high gas mileage ratings: the Hyundai Elantra (40 mpg) and the Toyota Prius (48 mpg). The Bentley Mulsanne stands out for its low gas mileage rating (18 mpg). All three of these values seem like clear outliers.

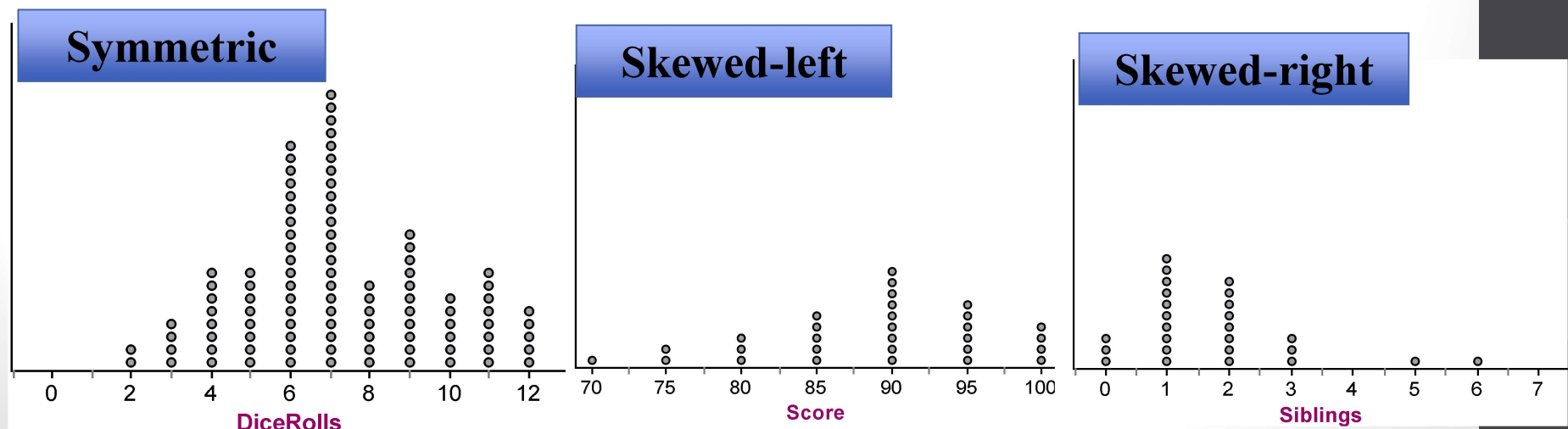
Describing Shape

When you describe a distribution's shape, concentrate on the main features. Look for rough **symmetry** or clear **skewness**.

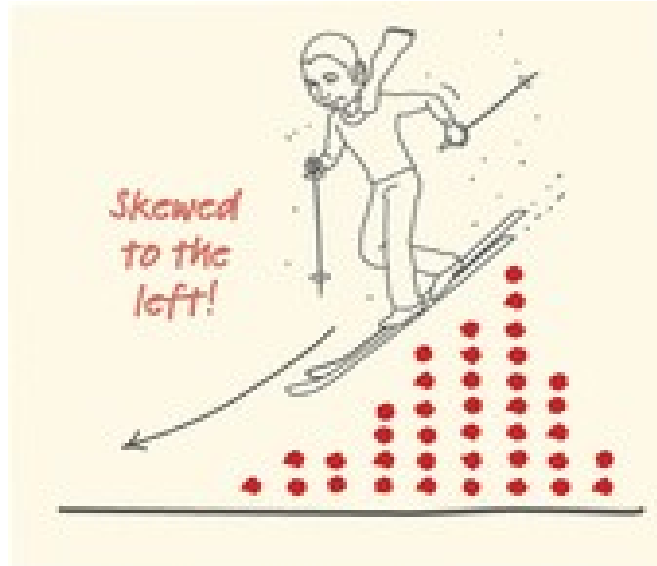
A distribution is roughly **symmetric** if the right and left sides of the graph are approximately mirror images of each other.

A distribution is **skewed to the right** (right-skewed) if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.

It is **skewed to the left** (left-skewed) if the left side of the graph is much longer than the right side.



The direction of skewness is the direction of the long tail, not the direction where most observations are clustered. See the drawing below for a cute but corny way to help you keep this straight.



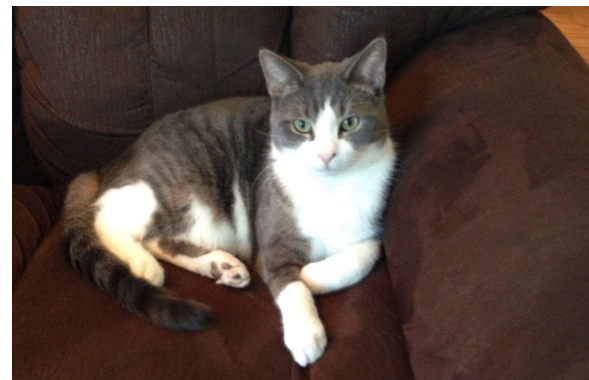
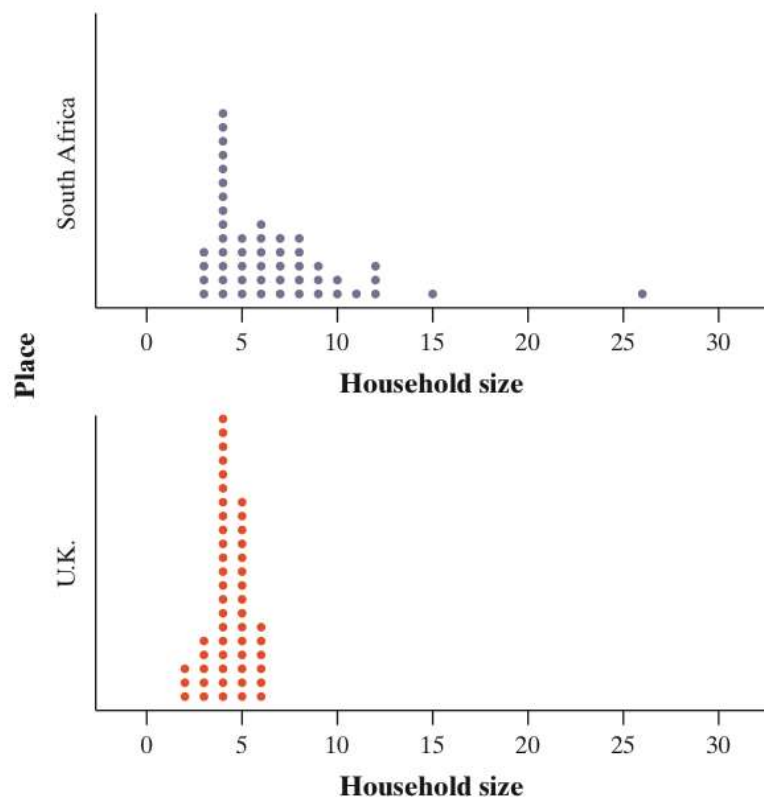
For his own safety, which way should Mr. Starnes go “skewing”?

Comparing Distributions

Some of the most interesting statistics questions involve comparing two or more groups.

Always discuss shape, center, spread, and possible outliers whenever you compare distributions of a quantitative variable.

Example 3: How do the numbers of people living in households in the United Kingdom (U.K.) and South Africa compare? To help answer this question, we used CensusAtSchool's "Random Data Selector" to choose 50 students from each country. The figure below is a dotplot of the household sizes reported by the survey respondents. Compare the distributions of household size for these two countries.



Shape: The distribution of household size for the U.K. sample is roughly symmetric and unimodal, while the distribution for the South Africa sample is skewed to the right and unimodal.

Center: Household sizes for the South African students tended to be larger than for the U.K. students. The midpoints of the household sizes for the two groups are 6 people and 4 people, respectively.

Spread: The household sizes for the South African students vary more (from 3 to 26 people) than for the U.K. students (from 2 to 6 people).

Outliers: There don't appear to be any outliers in the U.K. distribution. The South African distribution seems to have two outliers in the right tail of the distribution—students who reported living in households with 15 and 26 people.

AP EXAM TIP: When comparing distributions of quantitative data, it's not enough just to list values for the center and spread of each distribution. You have to explicitly *compare* these values, using words like “greater than,” “less than,” or “about the same as.”

Stem plots

Another simple graphical display for small data sets is a **stem plot**. (Also called a **stem-and-leaf plot**.) Stemplots give us a quick picture of the distribution while including the actual numerical values.

How to make a stem plot:

- *Separate each observation into a **stem** (all but the final digit) and a **leaf** (the final digit).
- *Write all possible stems from the smallest to the largest in a vertical column and draw a vertical line to the right of the column.
- *Write each leaf in the row to the right of its stem.
- *Arrange the leaves in increasing order out from the stem.
- *Provide a key that explains in context what the stems and leaves represent.

Example 4: How many pairs of shoes does a typical teenager have? To find out, a group of AP[®] Statistics students conducted a survey. They selected a random sample of 20 female students from their school. Then they recorded the number of pairs of shoes that each respondent reported having. Here are the data:

50	26	26	31	57	19	24	22	23	38
13	50	13	34	23	30	49	13	15	51

Create a stem plot of the data.

1 |
2 |
3 |
4 |
5 |

Stems

1 | 93335
2 | 664233
3 | 1840
4 | 9
5 | 0701

Add leaves

1 | 33359
2 | 233466
3 | 0148
4 | 9
5 | 0017

Order leaves

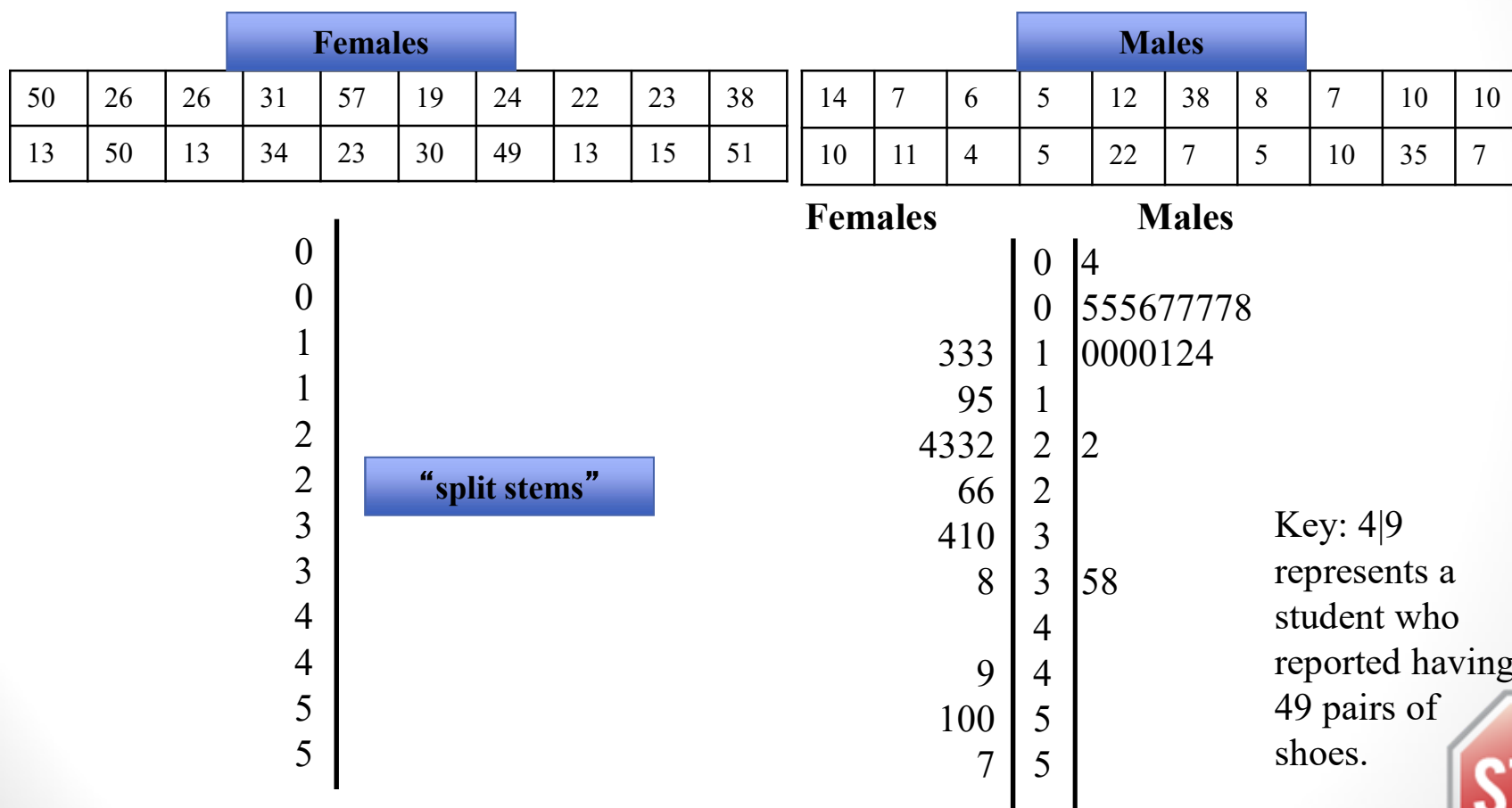
Key: 4|9
represents a
female student
who reported
having 49
pairs of shoes.

Add a key

Splitting Stems and Back-to-Back Stem plots

*When data values are “bunched up”, we can get a better picture of the distribution by **splitting stems**.

*Two distributions of the same quantitative variable can be compared using a **back-to-back stem plot** with common stems.



Histograms

*Quantitative variables often take many values. A graph of the distribution may be clearer if nearby values are grouped together.

*The most common graph of the distribution of one quantitative variable is a **histogram**.

How to make a histogram:

*Divide the range of data into classes of equal width.

*Find the count (frequency) or percent (relative frequency) of individuals in each class.

*Label and scale your axes and draw the histogram. The height of the bar equals its frequency. Adjacent bars should touch, unless a class contains no individuals.

Example 5: What percent of your home state's residents were born outside the United States? A few years ago, the country as a whole had 12.5% foreign-born residents, but the states varied from 1.2% in West Virginia to 27.2% in California. The following table presents the data for all 50 states.

The *individuals* in this data set are the states. The *variable* is the percent of a state's residents who are foreign-born. It's much easier to see from a graph than from the table how your state compared with other states.

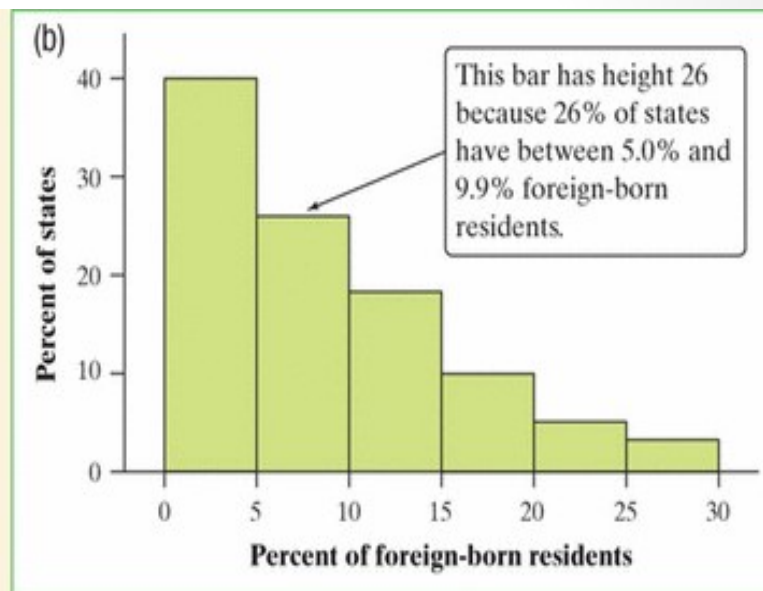
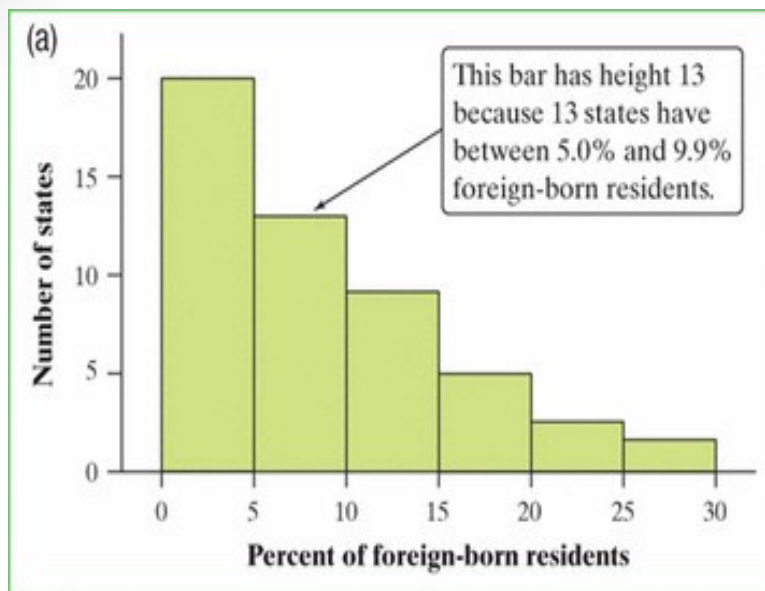
State	Percent	State	Percent	State	Percent
Alabama	2.8	Louisiana	2.9	Ohio	3.6
Alaska	7.0	Maine	3.2	Oklahoma	4.9
Arizona	15.1	Maryland	12.2	Oregon	9.7
Arkansas	3.8	Massachusetts	14.1	Pennsylvania	5.1
California	27.2	Michigan	5.9	Rhode Island	12.6
Colorado	10.3	Minnesota	6.6	South Carolina	4.1
Connecticut	12.9	Mississippi	1.8	South Dakota	2.2
Delaware	8.1	Missouri	3.3	Tennessee	3.9
Florida	18.9	Montana	1.9	Texas	15.9
Georgia	9.2	Nebraska	5.6	Utah	8.3
Hawaii	16.3	Nevada	19.1	Vermont	3.9
Idaho	5.6	New Hampshire	5.4	Virginia	10.1
Illinois	13.8	New Jersey	20.1	Washington	12.4
Indiana	4.2	New Mexico	10.1	West Virginia	1.2
Iowa	3.8	New York	21.6	Wisconsin	4.4
Kansas	6.3	North Carolina	6.9	Wyoming	2.7
Kentucky	2.7	North Dakota	2.1		

Here is a frequency table and a relative frequency table for these data:

Frequency table	
Class	Count
0 to < 5	20
5 to < 10	13
10 to < 15	9
15 to < 20	5
20 to < 25	2
25 to < 30	1
Total	50

Relative frequency table	
Class	Percent
0 to < 5	40
5 to < 10	26
10 to < 15	18
15 to < 20	10
20 to < 25	4
25 to < 30	2
Total	100

Notice that the frequencies add to 50, the number of individuals (states) in the data, and that the relative frequencies add to 100%.



AP EXAM TIP: If you're asked to make a graph on a free-response question, be sure to label and scale your axes. Unless your calculator shows labels and scaling, don't just transfer a calculator screen shot to your paper.

Using Histograms Wisely

Here are several cautions based on common mistakes students make when using histograms.

Cautions!

- *Don't confuse *histograms* and *bar graphs*.
- *Don't use counts (in a frequency table) or percents (in a relative frequency table) as data.
- *Use percents instead of counts on the vertical axis when comparing distributions with different numbers of observations.
- *Just because a graph looks nice, it's not necessarily a meaningful display of data.