1.2.1 Dotplots

Dotplot - A simple graph that shows each data value as a dot above its location on a number line

Example – Gooooaaaaalllllll!
How to make a dotplot

How good was the 2004 U.S. women's soccer team? With players like Brandi Chastain, Mia Hamm, and Briana Scurry, the team put on an impressive showing en route to winning the gold medal at the 2004 Olympics in Athens. Here are data on the number of goals scored by the team in 34 games played during the 2004 season:

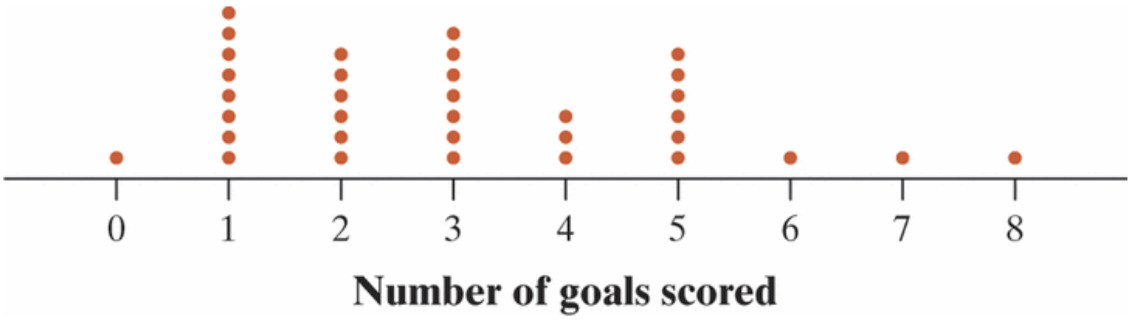
3	0	2	7	8	2	4	3	5	1	1	4	5	3	1	1	3
3	3	2	1	2	2	2	4	3	5	6	1	5	5	1	1	5

Here are the steps in making a dotplot:

1. *Draw a horizontal axis (a number line) and label it with the variable name*. In this case, the variable is number of goals scored.

2. *Scale the axis*. Start by looking at the minimum and maximum values of the variable. For these data, the minimum number of goals scored was 0, and the maximum was 8. So we mark our scale from 0 to 8, with tick marks at every whole-number value.

3. *Mark a dot above the location on the horizontal axis corresponding to each data value*. The figure below displays a completed dotplot for the soccer data

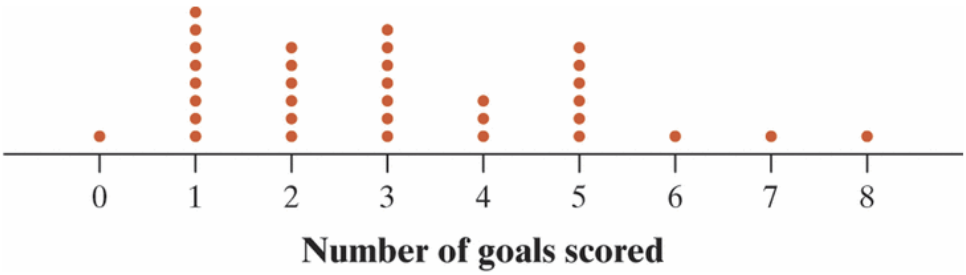


Making a graph is not an end in itself. The purpose of a graph is to help us understand the data. After you make a graph, always ask, "What do I see?"

How to Examine the Distribution of a Quantitative Variable

In any graph, look for the **overall pattern** and for striking **departures** from that pattern.

- You can describe the overall pattern of a distribution by its **shape, center**, and **spread**.
- An important kind of departure is an **outlier**, an individual value that falls outside the



Shape: The dotplot has a peak at 1. This indicates that the team's most frequent number of goals scored in games that season (known as the Mode) was 1. In most of its games, the U.S. women's soccer team scored between 1 and 5 goals. However, the distribution has a long tail to the right.

Center: We can describe the center by finding a value that divides the observations so that about half take larger values and about half take smaller values. This value is called the *median* of the distribution. In the figure above, the median is 3. That is, in a typical game during the 2004 season, the U.S. women's soccer team scored about 3 goals. Of course, we could also summarize the center of the distribution by calculating the average *(mean)* number of goals scored per game. For the 2004 season, the team's mean was 3.06 goals.

Spread: The spread of a distribution tells us how much *variability* there is in the data. One way to describe the variability is to give the smallest and largest values. The spread in the figure is from 0 goals to 8 goals scored. Alternatively, we can compute the range. Calculate the range of the distribution by subtracting the smallest value from the largest value. For the data above, the range is 8 − 0 = 8 goals.

Outliers: Was the game in which the women's team scored 8 goals an outlier? How about the team's 7-goal game? These values differ somewhat from the overall pattern. However, they don't clearly stand apart from the rest of the distribution. For now, let's agree to call attention only to potential outliers that suggest something special about an observation. In a later section, we'll establish a procedure for determining whether a particular data value is an outlier.

When describing a distribution of quantitative data, don't forget your SOCS (shape, outliers, center, spread)!

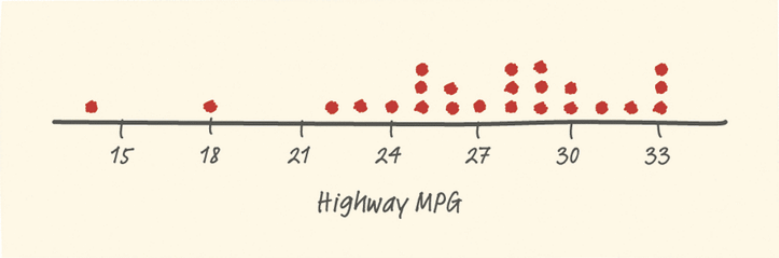
Example – Are Your Driving a Gas Guzzler?
Interpreting a Dotplot

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 below displays the EPA estimates of highway gas mileage in miles per gallon(mpg) for a sample of 24 model year 2009 midsize cars.

Model	Mpg	Model	Mpg	Model	Mpg
Acura RL	22	Dodge Avenger	30	Mercury Milan	29
Audi A6 Quattro	23	Hyundai Elantra	33	Mitsubishi Galant	27
Bentley Arnage	14	Jaguar XF	25	Nissan Maxima	26
BMW 528I	28	Kia Optima	32	Rolls Royce Phantom	18
Buick Lacrosse	28	Lexus GS 350	26	Saturn Aura	33
Cadillac CTS	25	Lincoln MKZ	28	Toyota Camry	31
Chevrolet Malibu	33	Mazda 6	29	Volkswagen Passat	29
Chrysler Sebring	30	Mercedes-Benz E350	24	Volvo S80	25

Here is a dotplot of the data:



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

1.2.2 Describing Shape

When you describe a distribution's shape, concentrate on the main features. Look for major peaks, not for minor ups and downs in the graph. Look for clusters of values and obvious gaps. Look for potential outliers, not just for the smallest and largest observations. Look for rough symmetry or clear skewness.

Symmetry - If the right and left sides of a graph are approximately mirror images of each other

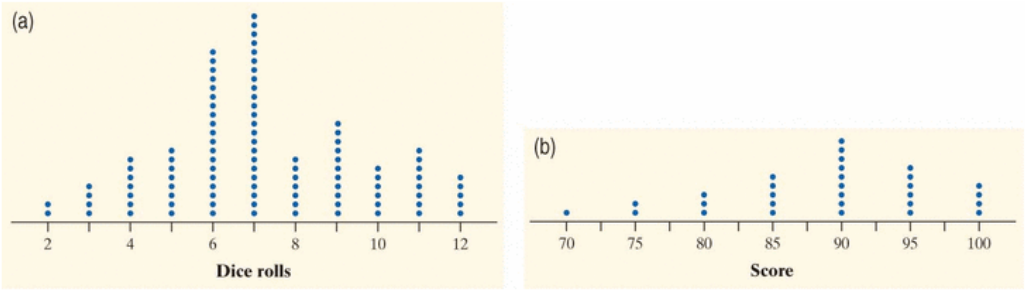
Skewness - A distribution is **skewed to the right** if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side. It is **skewed to the left** if the left side of the graph is much longer than the right side.

Skewed Right Distribution

Skewed Left Distribution

Example – Die Rolls and Quiz Scores
Describing Shape

The below figure displays dotplots for two different sets of quantitative data. Let's practice describing the shapes of these distributions. The dotplot to the left shows the results of rolling a pair of fair, six-sided dice and finding the sum of the up-faces 100 times. This distribution is roughly symmetric. The dotplot to the right shows the scores on an AP Statistics class's first quiz. This distribution is skewed to the left.



Although the dotplots in the previous example have different shapes, they do have something in common. Both are unimodal, that is, they have a single peak: the graph of dice rolls at 7 and the graph of quiz scores at 90. (We don't count minor ups and downs in a graph, like the "bumps" at 9 and 11 in the dice rolls dotplot, as "peaks.")

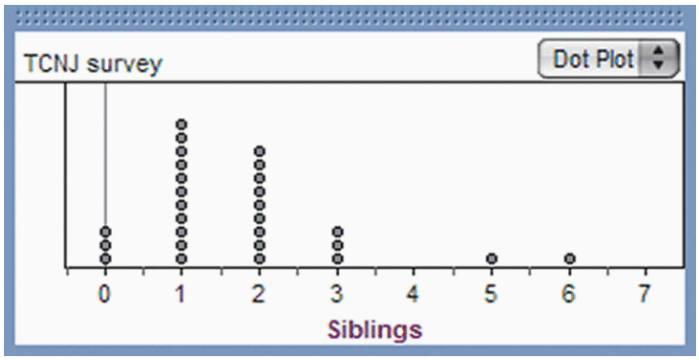
Unimodal - Describes a graph of quantitative data with a single peak.

Bimodal - Describes a graph of quantitative data with two clear peaks.

Multimodal - Describes a graph of quantitative data with more than two clear peaks.

Check Your Understanding

The Fathom dotplot displays data on the number of siblings reported by each student in a statistics class.



1. Describe the shape of the distribution

2. Describe the center of the distribution

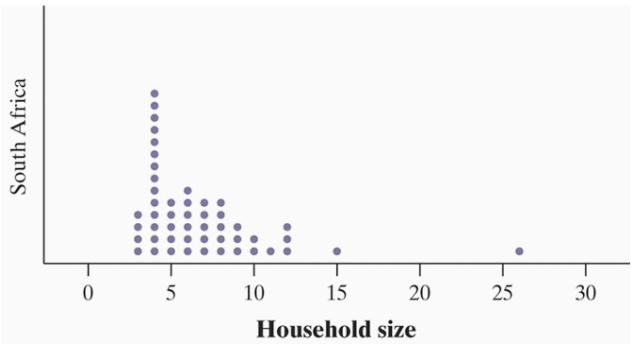
3. Describe the spread of the distribution

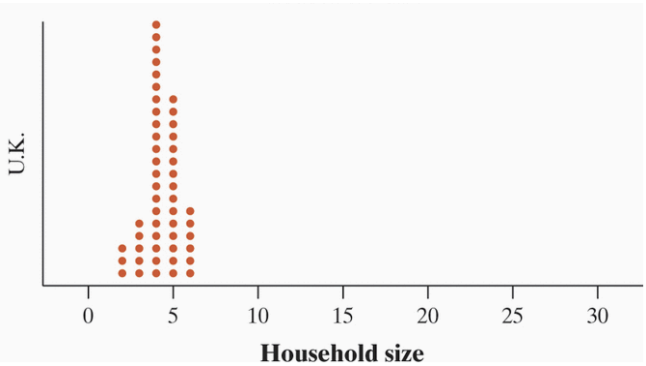
4. Identify any potential outliers

1.2.3 Comparing Distributions

Example – Household Size: U.K. versus South Africa
Comparing distributions

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 figures below is a dotplot of the household sizes reported by the survey respondents.





Compare the distributions of household size for these two countries (don't forget your SOCS!).

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 must explicitly *compare* these values, using words like "greater than," "less than," or "about the same as."

1.2.4 Stemplots

Stemplot - *(also called a stem-and-leaf plot)* A simple graphical display for fairly small data sets that gives a quick picture of the shape of a distribution while including the actual numerical values in the graph. Each observation is separated into a **stem**, consisting of all but the final digit, and a **leaf**, the final digit. The stems are arranged in a vertical column with the smallest at the top. Each leaf is written in the row to the right of its stem, with the leaves arranged in increasing order out from the stem

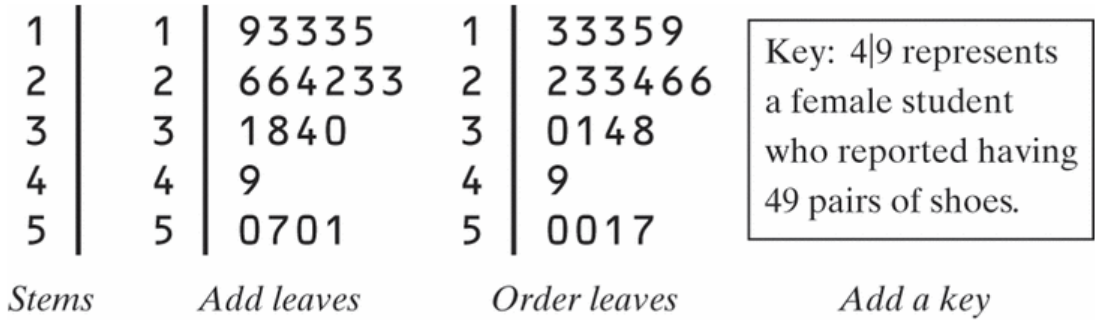
Example – How Many Shoes?
Making a Stemplot

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

Below are the steps for making a stemplot.

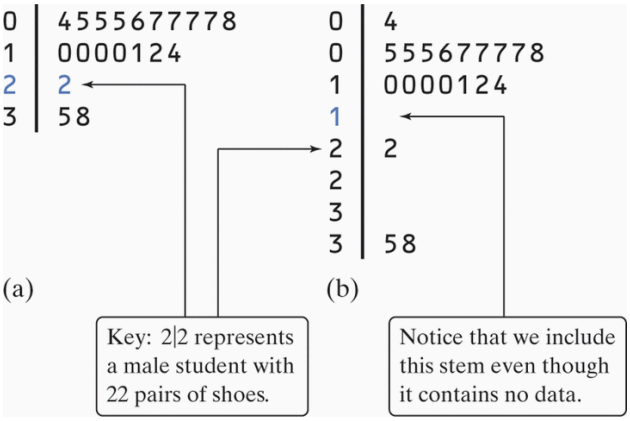
1. Separate each observation into a **stem,** consisting of all but the final digit, and a **leaf,** the final digit. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column. Do not skip any stems, even if there is no data value for a particular stem. For the data above, the tens digits are the stems, and the ones digits are the leaves. The stems run from 1 to 5.
2. *Write each leaf in the row to the right of its stem*. For example, the female student with 50 pairs of shoes would have stem 5 and leaf 0, while the student with 31 pairs of shoes would have stem 3 and leaf 1.
3. *Arrange the leaves in increasing order out from the stem*.
4. *Provide a key that explains in context what the stems and leaves represent*.



The AP Statistics students in the previous example also collected data from a random sample of 20 male students at their school. Below are the numbers of pairs of shoes reported by each male in the sample:

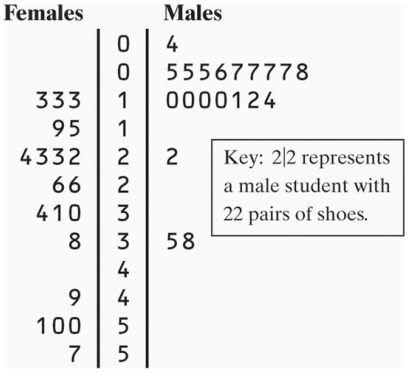
14	7	6	5	12	38	8	7	10	10
10	11	4	5	22	7	5	10	35	7

What would happen if we tried the same approach as before: using the first digits as stems and the last digits as leaves? The completed stemplot is shown in the figure below. What shape does this distribution have? It is difficult to tell with so few stems. We can get a better picture of male shoe ownership by **splitting stems**.



Splitting stems - A method for spreading out a stemplot that has too few stems.

What if we want to compare the number of pairs of shoes that males and females have? That calls for a **back-to-back stemplot** with common stems. The leaves on each side are ordered out from the common stem. The figure below is a back-to-back stemplot for the male and female shoe data. Note that we have used the split stems from the above figure as the common stems. The values on the right are the male data. The values on the left are the female data, ordered out from the stem from right to left.



Below are a few tips to consider when making stemplots:

- Stemplots do not work well for large data sets, where each stem must hold a large number of leaves.
- There is no magic number of stems to use, but five is a good minimum. Too few or too many stems will make it difficult to see the distribution's shape.
- If you split stems, be sure that each stem is assigned an equal number of possible leaf digits (two stems, each with five possible leaves; or five stems, each with two possible leaves).
- You can get more flexibility by rounding the data so that the final digit after rounding is suitable as a leaf. Do this when the data have too many digits. For example, in reporting teachers' salaries, using all five digits (for example, $42,549) would be unreasonable. It would be better to round to the nearest thousand and use 4 as a stem and 3 as a leaf.
- Instead of rounding, you can also *truncate* (remove one or more digits) when data have too many digits. The teacher's salary of $42,549 would truncate to $42,000.

Check Your Understanding

1. Use the back-to-back stemplot below to write a few sentences comparing the number of pairs of shoes owned by males and females. Be sure to address shape, center, spread, and outliers.

Females		Males
	0	4
	0	555677778
333	1	0000124
95	1	
4332	2	2
66	2	
410	3	
8	3	58
	4	
9	4	
100	5	
7	5	

Key: 2|2 represents a male student with 22 pairs of shoes.

Multiple Choice: Select the best answer for Questions 2 through 4.
Here is a stemplot of the percents of residents aged 65 and older in the 50 states and the District of Columbia. The stems are whole percents and the leaves are tenths of a percent.

6	8
7	
8	8
9	79
10	08
11	15566
12	012223444457888999
13	01233333444899
14	02666
15	23
16	8

Key: 8|8 represents a state in which 8.8% of residents are 65 and older.

2. The low outlier is Alaska. What percent of Alaska residents are 65 or older?

(a) 0.68 (b) 6.8 (c) 8.8 (d) 16.8 (e) 68

3. Ignoring the outlier, the shape of the distribution is

(a) skewed to the right (c) skewed to the left. (e) skewed to the middle.
(b) roughly symmetric (d) bimodal.

4. The center of the distribution is close to

(a) 13.3%. (b) 12.8%. (c) 12.0%. (d) 11.6%. (e) 6.8% to 16.8%.

1.2.5 Histograms

Quantitative variables often take many values. A graph of the distribution is clearer if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a **histogram**.

Example – Foreign-Born Residents
Making a histogram

What percent of your home state's residents were born outside the United States? The country as a whole has 12.5% foreign-born residents, but the states vary from 1.2% in West Virginia to 27.2% in California. The table below 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 compares 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		

Steps to make a histogram:

1. Divide the range of the data into classes of equal width. The data in the table vary from 1.2 to 27.2, so we might choose to use classes of width 5, beginning at 0:

0–5 5–10 10–15 15–20 20–25 25–30

But we need to specify the classes so that each individual falls into exactly one class. For instance, what if a state had exactly 5.0% of its residents born outside the United States? Since a value of 0.0% would go in the 0–5 class, we'll agree to place a value of 5.0% in the 5–10 class, a value of 10.0% in the 10–15 class, and so on. In reality, then, our classes for the percent of foreign-born residents in the states are

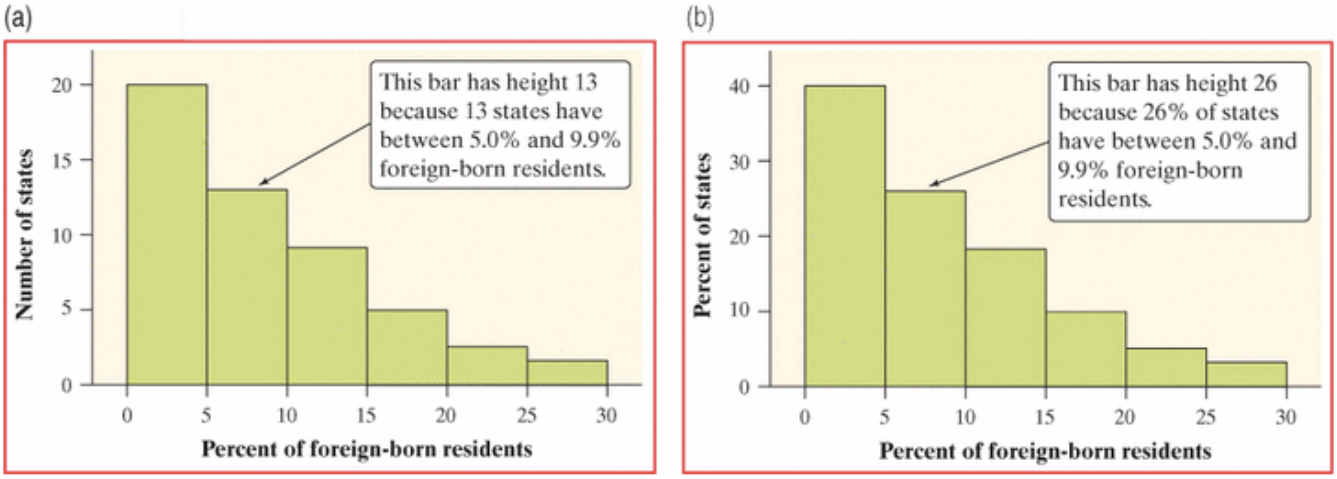
0 to <5 5 to <10 10 to <15 15 to <20 20 to <25 25 to <30

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

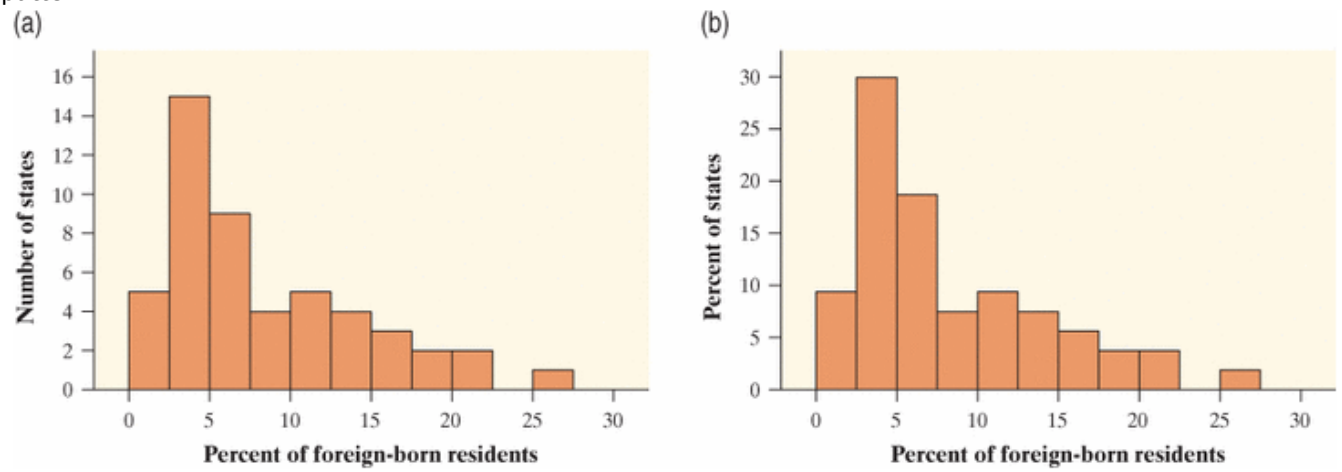
3. *Label and scale your axes and draw the histogram*. Label the horizontal axis with the variable whose distribution you are displaying. That's the percent of a state's residents who are foreign-born. The scale on the horizontal axis runs from 0 to 30 because that is the span of the classes we chose. The vertical axis contains the scale of counts or percents. Each bar represents a class. The base of the bar covers the class, and the bar height is the class frequency or relative frequency. Draw the bars with no horizontal space between them unless a class is empty, so that its bar has height zero.



Histogram (a) is a frequency histogram. Histogram (b) is relative frequency histogram.

Below is a description of the Histogram:
***Shape*:** The distribution is skewed to the right. A majority of states have fewer than 10% foreign-born residents, but several states have much higher percents, so that the graph extends quite far to the right of its peak. The distribution has a *single peak* at the left, which represents states in which between 0% and 4.9% of residents are foreign-born.
***Center*:** From the graph, we see that the midpoint (median) would fall somewhere in the 5.0% to 9.9% class. Remember that we're looking for the value having 25 states with smaller percents foreign-born and 25 with larger. (Arranging the observations from the table in order of size shows that the median is 6.1%.)
***Spread*:** The histogram shows that the percent of foreign-born residents in the states varies from less than 5% to over 25%. (Using the data in the table, we see that the range is 27.2% − 1.2% = 26.0%.)
***Outliers*:** We don't see any observations outside the overall single-peaked, right-skewed pattern of the distribution.

The histograms below shows (a) a frequency histogram and (b) a relative frequency histogram of the same distribution, with classes half as wide. The new classes are 0–2.4, 2.5–4.9, etc. Now California, at 27.2%, stands out as a potential outlier in the right tail. The choice of classes in a histogram can influence the appearance of a distribution. Histograms with more classes show more detail but may have a less clear pattern.



Check Your Understanding

Many people believe that the distribution of IQ scores follows a "bell curve," like the one shown in the margin. But is this really how such scores are distributed? The IQ scores of 60 fifth-grade students chosen at random from one school are shown below.

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

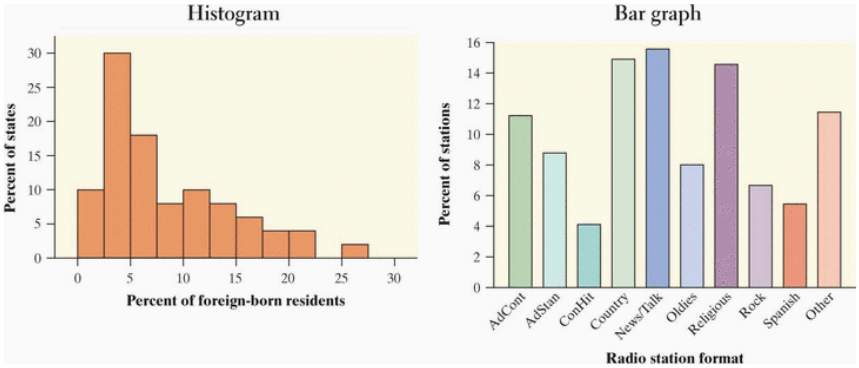
1. Construct a histogram that displays the distribution of IQ scores effectively.

2. Describe what you see. Is the distribution bell-shaped?

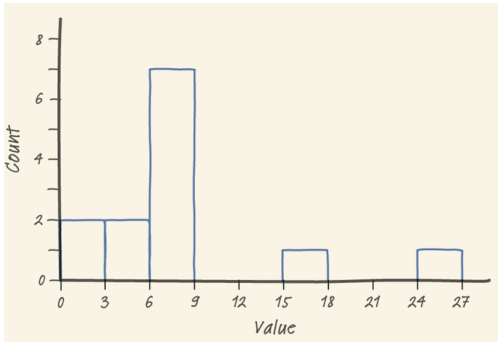
1.2.6 Using Histograms Wisely

Cautions

1. *Don't confuse histograms and bar graphs*. Although histograms resemble bar graphs, their details and uses are different. A histogram displays the distribution of a quantitative variable. The horizontal axis of a histogram is marked in the units of measurement for the variable. A bar graph is used to display the distribution of a categorical variable or to compare the sizes of different quantities. The horizontal axis of a bar graph identifies the categories or quantities being compared. Draw bar graphs with blank space between the bars to separate the items being compared. Draw histograms with no space, to show the equal-width classes. For comparison, here is one of each type of graph from previous examples.

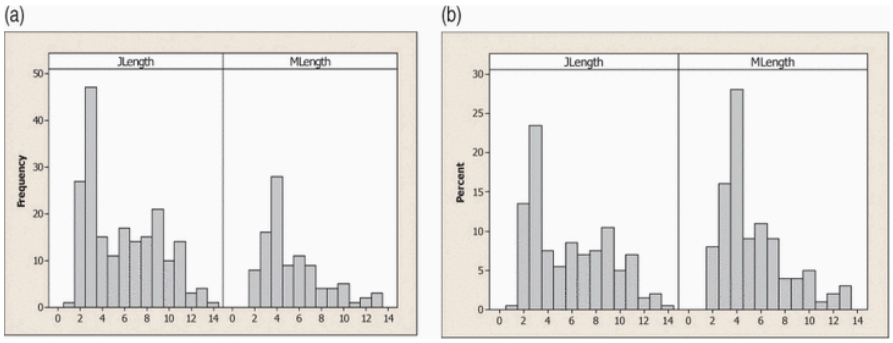


2. *Don't use counts (in a frequency table) or percents (in a relative frequency table) as data*. Below is a frequency table displaying the lengths (number of letters) of the first 100 words in a journal article. to display these data. Can you see what Billy did wrong? (He used the counts as data when drawing the histogram—so there were two counts of 1, two counts between 3 and 5, and so on.)

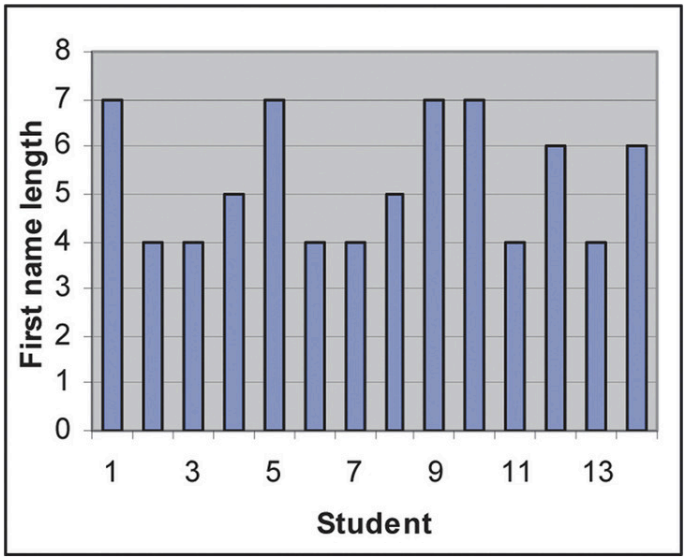


Length:	1	2	3	4	5	6	7	8	9	10	11	12	13
Count:	1	15	25	7	5	7	8	7	7	6	8	3	1

3. *Use percents instead of counts on the vertical axis when comparing distributions with different numbers of observations*. Mary was interested in comparing the reading levels of a medical journal and an airline's in-flight magazine. She counted the number of letters in the first 200 words of an article in the medical journal and of the first 100 words of an article in the airline magazine. Mary then used Minitab statistical software to produce the histograms shown in figure (a). This figure is misleading—it compares frequencies, but the two samples were of very different sizes (100 and 200). Using the same data, Mary's teacher produced the histograms in figure (b). By using relative frequencies, this figure provides an accurate comparison of word lengths in the two samples.

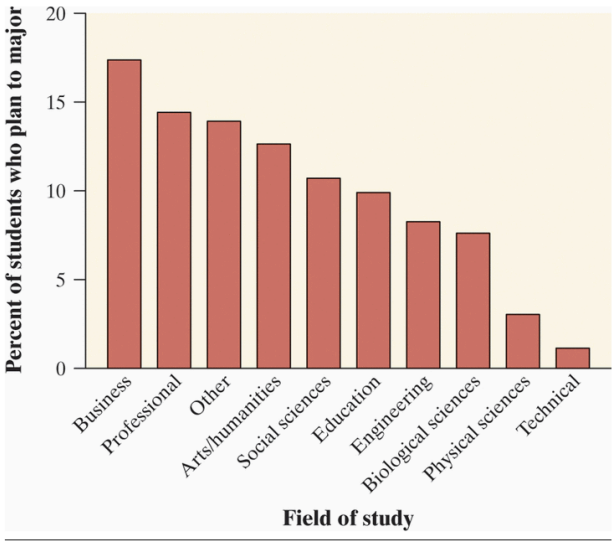


4. *Just because a graph looks nice, it's not necessarily a meaningful display of data*. The students in a small statistics class recorded the number of letters in their first names. One student entered the data into an Excel spreadsheet and then used Excel's "chart maker" to produce the graph shown. What kind of graph is this? It's neither a bar graph nor a histogram. Both of these types of graphs display the number or percent of individuals in a given category or class. This graph shows the individual data values, in the order that they were entered into the spreadsheet. It is not a very meaningful display of the data.



Check Your Understanding

Questions 1 and 2 relate to the following setting. About 1.6 million first-year students enroll in colleges and universities each year. What do they plan to study? The graph displays data on the percents of first-year students who plan to major in several discipline areas.



1. Is this a bar graph or a histogram? Explain.

2. Would it be correct to describe this distribution as right-skewed? Why or why not?