TOPIC D2: GRAPHING DATA

SPOTLIGHT: WHO ARE THE BASEBALL PLAYERS?

Baseball has been traditionally called the “great American game,” and has been played at a professional level in the United States since 1871. Although the rules of the game have remained virtually unchanged since then, there have been dramatic changes in the types of men who play the game. These changes are described in *The New Bill James Historical Baseball Abstract*. In the beginning years, professional baseball was played primarily by U.S. immigrants and residents of eastern U.S. cities such as Brooklyn, Philadelphia, and Baltimore. One prominent group of immigrants who played baseball was the Irish. Toward the end of the century, many players were men who had played baseball in college. The creation of the American League in 1901 introduced many players from the Midwest region of the U.S. At this time, the baseball rosters had players from many different vocations. In the 1920’s, more men from rural parts of the country joined professional baseball, and the number of players with college educations declined. The 1930’s introduced more players from Southern U.S. and California.

Before 1947, only white Caucasians played baseball in the Major Leagues, and African-American baseball players played in their own Negro League. Although Jackie Robinson broke the color barrier in 1947 when he was signed by the Brooklyn Dodgers, professional baseball was slow to adopt African-Americans and other minorities to their teams. But by the 1960’s, a large number of African-American and Latin-Americans played the game. The 1970’s saw the introduction of a suburban generation of American players who learned baseball at an early age through organized leagues; during this time the number of African-American players reached its peak. In recent years, professional baseball has seen a rise in the number of Latin-American players and American players of Latino ancestry, and a decline in the number of African-American players. Professional baseball has also recently seen the introduction of players from the Far East, especially Japan.

PREVIEW

In this topic, you begin your exploration of data. You saw in Topic D1 that variables differ by how they are measured. Now you'll come to see that the appropriate graph for a particular variable depends upon the measurement type. By graphing quantitative data, you will be introduced to the notion of a dataset's distribution, which illustrates the variability in the data values. You will learn about a distribution's shape and begin to make informal judgments about the center and spread of the values. You'll also gain experience interpreting different graphs.

In this topic your learning objectives are to:

- Understand how to construct and interpret different graphs for a single collection of data.
- Understand there are choices in constructing a graph and the appearance and the usefulness of a graph can be dependent of these choices.
- Use the graph to write a descriptive paragraph about the distribution of a dataset that contains information about the shape, center, spread, and any unusual characteristics of the distribution.
- Compare the usefulness of different types of graphs in representing a distribution.

NCTM Standards

✓In Grades 6-8, students should select, create, and use appropriate graphical representations of data.

✓In Grades 6-8, students should discuss and understand the correspondence between data sets and their graphical representations.

✓In Grades 9-12, all students should for, univariate measurement data, be able to display the distribution and describe its shape.

GRAPHING CATEGORICAL DATA

The opening spotlight describes the dramatic changes in the backgrounds of professional baseball players over the years. You may be asking yourself, “Who are the professional baseball players today?” To look more carefully at the backgrounds of current players, here are the countries of birth for each Major League Baseball player born in the year 1975. (These are the players that would be 30 years old in the 2005 baseball season. D.R. is the Dominican Republic and P.R. is Puerto Rico):

U.S.	U.S.	D.R.	U.S.	U.S.	U.S.	D.R.	U.S.	U.S.
D.R.	South Korea	U.S.	D.R.	U.S.	U.S.	D.R.	Panama	D.R.
U.S.	U.S.	U.S.	Cuba	D.R.	U.S.	U.S.	U.S.	U.S.
U.S.	U.S.	Venezuela	Cuba	Panama	U.S.	U.S.	Curacao	Venezuela
U.S.	U.S.	D.R.	Venezuela	U.S.	P.R.	U.S.	U.S.	Canada
U.S.	P.R.	U.S.	Venezuela	U.S.	U.S.	D.R.	D.R.	
U.S.	U.S.	D.R.	U.S.	U.S.	U.S.	U.S.	U.S.	
U.S.	U.S.	U.S.	U.S.	U.S.	Venezuela	U.S.	U.S.	
Venezuela	U.S.	U.S.	U.S.	U.S.	D.R.	U.S.	U.S.	
U.S.	U.S.	U.S.	U.S.	U.S.	U.S.	Venezuela	U.S.	
U.S.	P.R.	U.S.	U.S.	D.R.	U.S.	U.S.	U.S.	
U.S.	D.R.	U.S.	U.S.	Mexico	U.S.	Colombia	U.S.	
U.S.	U.S.	U.S.	Canada	D.R.	Venezuela	U.S.	Japan	
U.S.	U.S.	D.R.	U.S.	Nicaragua	Cuba	U.S.	D.R.	
U.S.	U.S.	D.R.	U.S.	U.S.	U.S.	D.R.	U.S.	
U.S.	U.S.	Mexico	U.S.	D.R.	Mexico	U.S.	Cuba	
P.R.	U.S.	Mexico	U.S.	Japan	Venezuela	U.S.	U.S.	
U.S.	P.R.	U.S.	U.S.	U.S.	Cuba	U.S.	P.R.	
U.S.	U.S.	U.S.	U.S.	U.S.	D.R.	U.S.	Venezuela	
U.S.	U.S.	Mexico	U.S.	U.S.	U.S.	U.S.	U.S.	
P.R.	U.S.	Venezuela	Canada	U.S.	U.S.	U.S.	U.S.	
U.S.	U.S.	Canada	U.S.	U.S.	U.S.	U.S.	U.S.	
U.S.	Australia	U.S.	U.S.	U.S.	U.S.	U.S.	U.S.	
D.R.	U.S.	Venezuela	U.S.	Panama	D.R.	Venezuela	U.S.	
U.S.	U.S.	D.R.	U.S.	D.R.	D.R.	U.S.	U.S.	

Display D2.1: Countries of birth for all Major League Baseball players born in 1975. (Source: Lahman baseball database, baseball1.com.)

A first step in organizing these categorical data is to construct a **frequency table** where you list the possible countries of origin and the corresponding counts or **frequencies** of each country. Looking at the frequency table below (Display D2.2), you see that a large number of players (135) are from the United States, but many other countries are represented in Major League Baseball. You may recognize that many of these countries are from the Latin America region. Because there are many countries with small counts, it is helpful to collapse the countries into the three categories "U.S.," "Latin America" and "Other." A frequency table using these new categories is also shown below (Display D2.3).

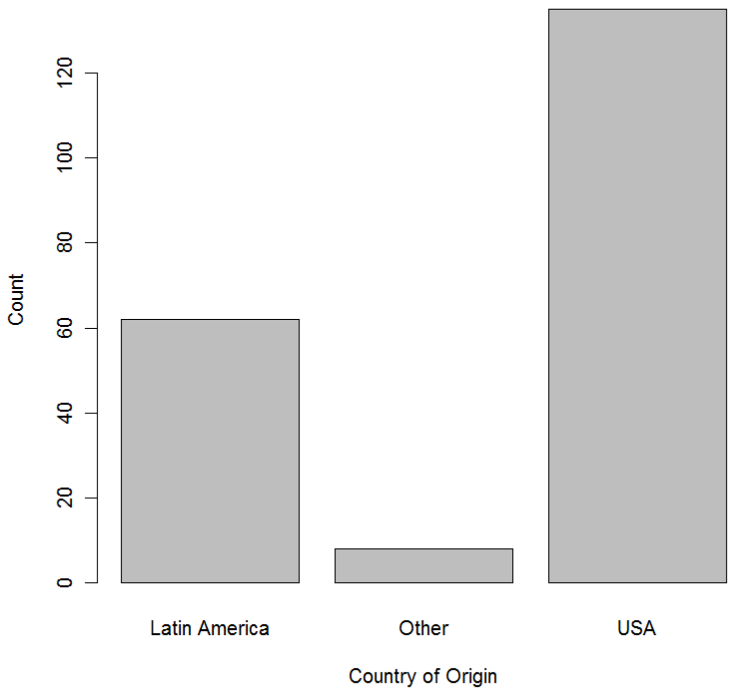
Country of Birth	**Frequency**
Australia	1
Canada	4
Columbia	1
Cuba	5
Curacao	1
D.R.	26
Japan	2
Mexico	5
Nicaragua	1
P.R.	7
Panama	3
South Korea	1
U.S.	135
Venezuela	13

Display D2.2: Frequencies of countries of birth for all Major League Baseball players born in 1975.

Country of Birth	**Frequency**
U.S.	135
Latin America	62
Other	8

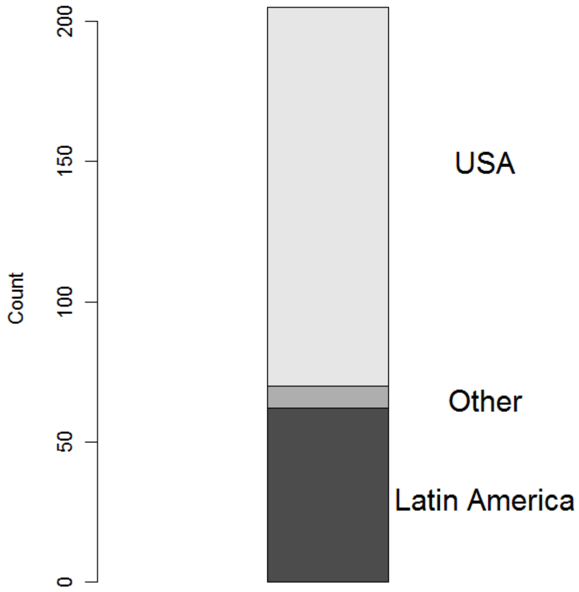
Display D2.3: Frequencies of countries of birth, by categories, for all Major League Baseball players born in 1975.

After you construct a frequency table, you are typically interested in describing the relative sizes of the category frequencies, and you can do this by constructing a suitable graph. There are several types of "good" graphs for representing categorical data –a *bar chart*, a *segmented bar chart*, and a *pie chart* are illustrated in this section. To construct a **bar chart,** you list the possible categories on one axis and then construct a bar for each category along the second axis, where the height (or length) of the bar corresponds to the frequency.

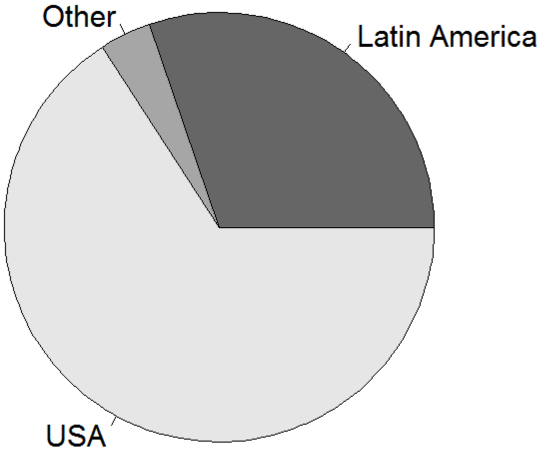


Display D2.4: Bar chart of countries of birth for all Major League Baseball players born in 1975.

A **segmented bar chart** is similar to a bar chart, but you stack the bars for the different categories into a single bar. The height (or length) of each section of the combined bar is proportional to the corresponding category frequency. You will see in Topic D4 that segmented bar charts are useful for comparing two or more batches of categorical data. A **pie chart** represents the entire group of ballplayers as a circle, and each country is represented by a slice of the pie. For a pie chart, the areas of the slices are proportional to the frequencies of the categories.



Display D2.5: Segmented bar chart of countries of birth for all Major League Baseball players born in 1975.



Display D2.6: Pie chart of countries of birth for all Major League Baseball players born in 1975.

This data exploration can be put in the context of the four basic components of statistics described in Topic D1. You began by asking "Who are the current baseball players?" To help answer this question, the places of birth of current major league baseball players born in the year 1975 were collected. These places of birth were grouped by country and the relative frequencies of countries were displayed using different types of graphs. By looking at any one of these three graphs, it is possible to draw a general conclusion about the nationality of current ballplayers. It is clear from the graphs that a majority of the ballplayers are born in the U.S. and most of the non-American-born players are from Latin America.

What is the best graph for representing categorical data? It is important that a graph present an accurate representation of the data in the frequency table. To be an accurate display, a graph should obey the **area principle,** where the area of the bar or object corresponding to a particular category is proportional to its frequency. All of the graphs above obey the area principle. (Recall that you saw several misleading graphs for categorical data that did *not* obey the area principle in Topic D1.)

Although all three graphs are accurate representations of the data, pie charts will not be featured in this book. A single pie chart helps you see the relative sizes of the counts for a single batch of categorical data. However, when several pie charts are used, it becomes difficult to compare batches of data because you have to visually compare the sizes of angles of the slices of the pie chart. It is generally easier for people to make visual comparisons of lengths of lines, and so bar charts and stacked bar charts are more effective for graphical comparison of batches. (The use of these graphs will be revisited in topic D4.)

PRACTICE: GRAPHING CATEGORICAL DATA

Suppose you are interested in buying a used sedan car. To help understand what types of cars are available under a cost of $10,000, you visit the website *usedcars.com* and it gives you a list of 50 available sedans that are located within 30 miles of your hometown. The table below gives the details of each car. (For mileage, under 70,000 miles is "low," 70,000–100,000 miles is "medium," and over 100,000 miles is "high.")

Model	**Year**	**Color**	**Mileage**
Chevrolet	1996	blue	high
Mercury	1999	white	high
Chevrolet	1997	blue	medium
Ford	1997	tan	high

Model	**Year**	**Color**	**Mileage**
Kia	2002	silver	low
Chrysler	1999	white	medium
Dodge	2000	blue	medium
Oldsmobile	1999	green	low

Plymouth	1997	burgundy	medium
Mercury	1997	red	high
Plymouth	1998	green	medium
Dodge	1999	red	low
Ford	1999	white	medium
Mercury	1997	blue	high
Ford	1996	copper	medium
Mitsubishi	2001	green	low
Ford	2000	white	low
Buick	1997	white	high
Pontiac	1998	green	medium
Mercury	1999	blue	medium
Buick	1999	green	medium
Ford	2000	green	low
Cadillac	1997	green	medium
Buick	2000	green	low
Chevrolet	2000	brown	low
Oldsmobile	2000	green	medium
Pontiac	2000	white	medium
Kia	2002	maroon	low
Dodge	2001	tan	medium
Buick	2000	maroon	medium
Pontiac	2000	silver	low
Ford	2001	blue	medium
Ford	2002	gray	medium
Chevrolet	2001	black	low
Oldsmobile	2000	green	medium
Chevrolet	2000	beige	low
Dodge	2002	brown	low
Ford	2002	red	low
Chevrolet	2000	beige	medium
Honda	1999	green	medium
Mercury	1999	tan	low
Mercury	1998	silver	medium
Mitsubishi	2002	green	low
Ford	2001	maroon	low
Saturn	2002	black	low
Pontiac	2002	burgundy	low
Nissan	2001	white	low
Ford	2002	gold	low
Honda	1999	silver	medium
Ford	2003	tan	low

Display D2.7: Details of 50 used sedan cars. (Source: *usedcars.com*)

1. Suppose you are interested in the manufacturers of these "cheap" used cars. So you classify the car models into four groups: those manufactured by General Motors (Cadillac, Chevrolet, Oldsmobile, Pontiac, Saturn and Buick), Ford (Ford and Mercury), Chrysler (Chrysler, Dodge, and Plymouth), and foreign (all others). Construct a frequency table and bar chart for the car model.

2. REFLECTION Based on your bar chart, make some comments about what you have learned about the models of these 50 cars. (For example, what are the popular and unpopular car manufacturers among these used cars?) Can you explain why there are so few foreign cars listed?

3. Suppose you consider the model year to be a categorical variable. Construct a segmented bar chart and a pie chart for this variable.

4. Construct a graph of the colors of the cars and use this graph to describe the popular and unpopular colors in this group of used cars.

5. REFLECTION If you were to purchase a used car, what other variables would you collect besides the ones listed above? Explain why these additional variables would be important to you.

GRAPHING QUANTITATIVE DATA – DISTRIBUTION AND SHAPE

In topic D1, you were introduced to a dataset from the U.S. Census Bureau that contained the percentages of adults completing high school and college for all states and the District of Columbia. (See Display D1.1 on page XXX.) Here are all of the data for the high school completion rates.

State	Completed High School (Percent)	State	Completed High School (Percent)	State	Completed High School (Percent)
Alabama	81.9	Kentucky	81.3	North Dakota	89.6
Alaska	91.6	Louisiana	81.2	Ohio	87.6
Arizona	83.8	Maine	89.7	Oklahoma	85.5
Arkansas	82	Maryland	88	Oregon	88.6
California	80.2	Massachusetts	88.7	Pennsylvania	87.5
Colorado	88.9	Michigan	88.1	Rhode Island	83.7
Connecticut	88.6	Minnesota	91.6	South Carolina	83.2
Delaware	87.2	Mississippi	79.9	South Dakota	90.3
District of Columbia	85.8	Missouri	86.5	Tennessee	83
Florida	85.2	Montana	90.9	Texas	79.6
Georgia	83.9	Nebraska	90.1	Utah	90.4
Hawaii	90.3	Nevada	83.5	Vermont	90.6
Idaho	87.9	New Hampshire	90.9	Virginia	85.9
Illinois	85.9	New Jersey	87.4	Washington	89.6
Indiana	86.2	New Mexico	82.4	West Virginia	82.2
Iowa	90.3	New York	84.1	Wisconsin	89.6
Kansas	89.5	North Carolina	83.6	Wyoming	91.7

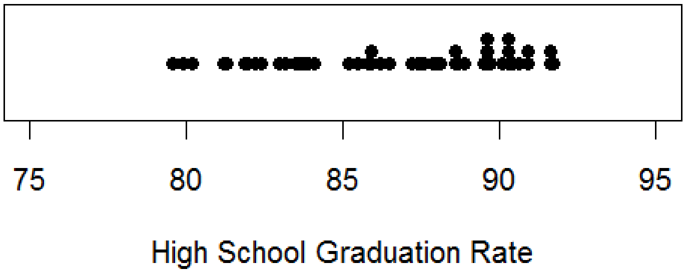
Display D2.8: The percentages of adults (25 years or over) that have completed high school, by state. (Source: U.S. Census Bureau.)

You may recall from Topic D1 that these completion rates were collected in order to answer questions similar to the following:

- What is a typical high school completion rate for the states?
- Are there sizable differences in the completion rates for states? Can states with high and low rates be identified?
- Are there particular states that stand out (either in the high end or in the low end) in terms of getting their high school students to graduate?

The first step in exploring these completion rates and answering the questions is to construct a graph. For quantitative data, a graph can help illuminate patterns—such as frequently occurring values, clusters, and gaps—that are not easily visible from a table of values. In a sense, a graph gives you a way to see how the variable actually varies.

A simple graph that is easy to construct by hand is a **dotplot** (often called a *line plot* in the school curriculum). You draw a number line covering the smallest and largest data values and then you place a dot on the number line for each data value. Here is a dotplot of the high school completion rates. This dotplot was created with *Fathom*, a software package designed to collect, explore, and analyze data. Note that the dots are placed on a higher level when they start to overlap. For example, there is a dot placed on a second level around 86 percent and the three dots corresponding to states with graduation rates of 90.3 percent are placed on three levels.



Display D2.9: Dotplot of high school completion rates.

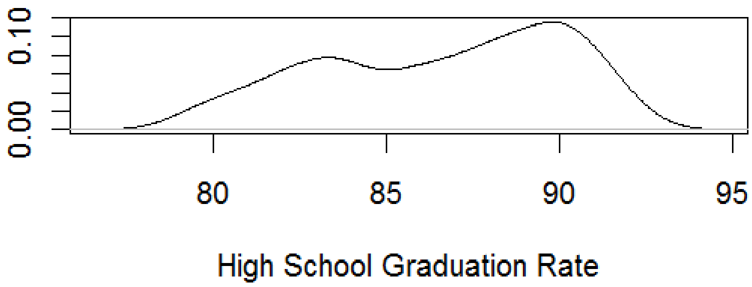
From the graph alone, you may begin to see answers to the questions at hand. For example, there are fairly sizeable differences—ranging between 79.6 and 91.7 percent—

in the high school completion rates. And because so many values are clustered between 87 and 92 percent, a typical completion rate might be somewhere in this interval.

What you see from the dotplot are the values that the variable takes and how often each occurs; this is called the **distribution** of high school completion rates. What do you look for in this distribution?

SHAPE

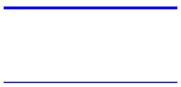
For one thing, you look at the general **shape** of the data. To help understand the shape of the data, you can draw a smooth curve over the dots in the dotplot. The curve helps you focus on the general pattern of the data rather than small gaps and clusters amongst the individual values.



Display D2.10: Smooth curve approximating the shape of the distribution of high school graduation rates.

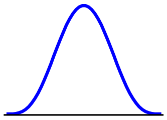
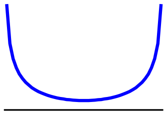
There are several common shapes that you may see in the smooth curve that you draw over the dotplot.

Uniform



In a **uniform** distribution, each data value occurs at roughly the same frequency. The overall shape of the distribution is flat.

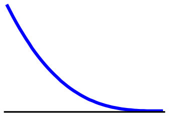
Symmetric

 or 

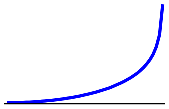
Symmetric describes a distribution in which the data values drop off (or increase) at the same rate at the left and right ends. You can imagine dividing the data into two halves, where the left half is approximately a mirror image of the right.

The symmetric distribution illustrated on the left—when the majority of the data values are in the middle—is frequently described as *bell-shaped* in school textbooks.

Skewed

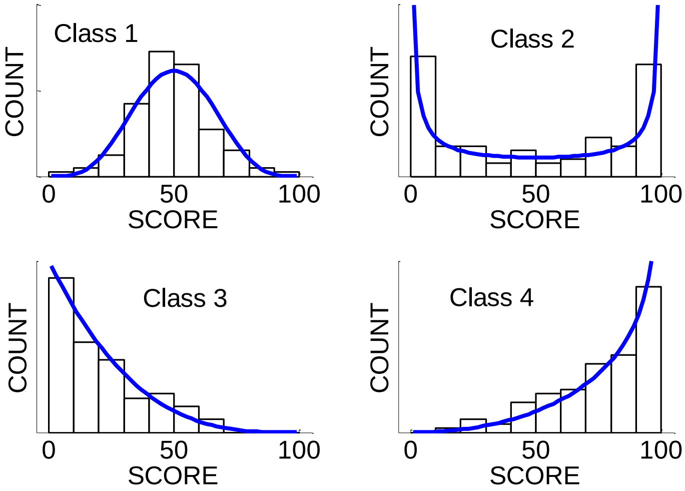


A distribution is called **skewed right** when the data values pile up at the low end and the frequency of data values decrease slowly as you move right toward larger values.



Conversely, **skewed left** is when the data values pile up for large values and the frequencies decrease slowly as you move left toward smaller values.

To help understand different shapes, consider these graphs of four batches of test scores corresponding to some hypothetical classes (Display D2.11). The shapes of the datasets are indicated by smooth curves drawn over the graphs. The scores of Class 1 have a symmetric shape with the frequencies of students scoring less than 50 dropping off in the same way as the frequencies of students scoring more than 50. Class 2's scores have a symmetric "u-shaped" distribution. For this class it was common to score close to 100 or close to 0. The scores of Class 3 are skewed right where low scores were the most common, and Class 4 scores are skewed left with a large number of high scores. You would probably be happiest with the test scores of Class 4, although the shape of the test scores for Class 1 is customary for standardized tests.



Display D2.11: Distributions of test scores for four hypothetical classes.

For the high school completion data, you see from the dotplot in Display D2.10 that there is a tight clump of states with high completion rates between 88 and 92 percent and there is a wide interval of states with low completion rates between 78 and 84 percent. So, the shape of the high school completion rates is somewhat skewed left.

CENTER

When examining a distribution, you also look for a **center,** or representative data value. You will learn more precise definitions in later topics, but for now look for a representative value that is located in the center of the distribution.

By looking at the dotplot in Display D2.10, you see that the center of the entire distribution is about 87. So, it might be reasonable to say that 87 percent is a representative high school completion rate for a state.

SPREAD

Along with noticing a central value, you want to say something about the **spread** of the percentages. The spread helps illuminate characteristics of the data that the center alone cannot. For example, consider two hypothetical classes of three students that take the same test. One class scores 40, 70, and 100, and the other class scores 69, 70, and 71. While both classes have a center of 70—and would appear to be identical if you were told only the center—describing the spread of the values would emphasize how drastically different the two classes performed.

Like the center, there are different ways of defining spread, and you will learn precise definitions later. One quick way to describe the spread is to name the interval that contains all of the values. From inspecting the data table in Display D2.8, you see that the minimum and maximum rates are 79.6 and 91.7 percent, respectively. So, the high school completion rates for all states and the District of Columbia fall in the interval [79.6, 91.7].

The interval that contains all of the data might be wide if there are one or more unusually small or large values. As an alternative, you can give an interval that contains a particular high percentage of the data. Looking at the dotplot in Display D2.9, you see that a large number of the values fall between 85 and 91 percent. Actually, 31 states have a high school completion rate greater than or equal to 85 percent but less than 91 percent. Because there are 51 states and 31/51 = 0.61, you may further describe the spread by saying, “About 61% of the states have high school completion rates between 85 and 91 percent.”

INTERESTING FEATURES

In addition to talking about the shape, center, and spread of the data, you also want to point out any interesting features of the distribution. These could include

- unusually small or large data values that stand out from the majority of the data; these are commonly called **outliers**
- gaps in the data
- clusters of several data values

For the high school completion rate data, no unusually small or large rates stand out. But it is interesting to note that the two smallest rates (79.6 and 79.9) correspond respectively to Texas and Mississippi, two southern states, and the three highest rates (91.6, 91.6 and 91.7) correspond to Alaska, Minnesota, and Wyoming, three northern states. This suggests that there might be a relationship between a state’s high school completion rate and its location. You’ll explore this relationship further in a later topic.

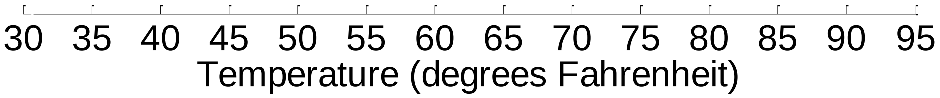
PRACTICE: GRAPHING QUANTITATIVE DATA

Here are the daily high temperatures (in degrees Fahrenheit) for the city of Atlanta, Georgia, in the month of March 2005.

March 2005						
		41	48	55	61	66
61	67	55	47	54	60	75
76	62	59	50	40	55	58
65	66	58	67	70	79	80
67	64	74	77	65		

Display D2.12: Daily high temperatures for Atlanta, Georgia. (Source: Georgia Automated Environmental Monitoring Network http://www.griffin.uga.edu/aemn/cgi-bin/AEMN.pl?site=GAAA&report=hi)

1. Construct a dotplot of the temperatures on the scale below.



2. Draw a smooth curve over the dotplot and describe the shape of the distribution of temperatures.

3. Give a center value for the temperatures and describe the spread by giving an interval that contains a high percentage of the data.

4. Are there any interesting features about these data, such as outliers, gaps, or clusters?

5. REFLECTION Is there any possible explanation for the large spread in temperatures in the data? (Think about how the weather in Atlanta changes in March.)

6. REFLECTION If you collected and graphed the daily high temperatures for Miami, Florida, in March 2005 and compared the distribution to the one above for Atlanta, would you expect to see any differences in the shape, center, and spread? How do you think the distribution of daily high temperatures in March 2005 for Minneapolis, Minnesota, would compare? Explain.

STEMPLOT

The dotplot is one way of graphing a single batch of quantitative data. There are alternative graphs that may also be useful. A **stemplot** is a clever way of quickly organizing a batch of data by hand using the digits of the data values. (Some school textbooks use the longer term *stem-and-leaf plot*.)

To construct a stemplot of the high school completion rates, you take each data value and divide it into two parts, called the **stem** and the **leaf**. For example, Alabama's rate of 81.9 percent could be written as a whole-number stem and a decimal leaf. A vertical line separates the two parts.

81 | 9

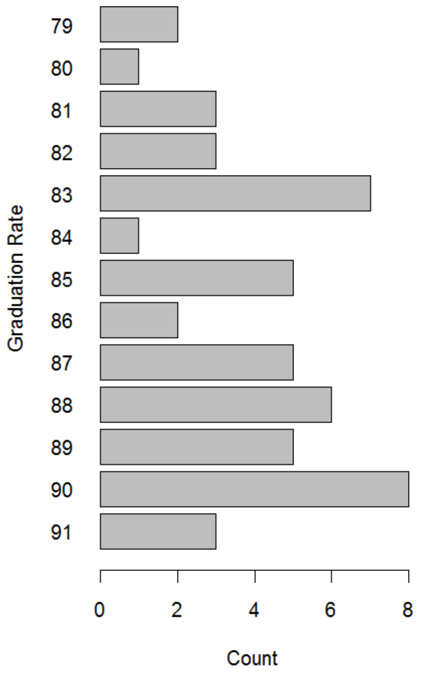
To create the entire stemplot, you first write down a chronological list of all possible stems on the left (here 79 to 91); you draw a vertical line to the right of the stems; and then you record each leaf after the appropriate stem. Doing this for all of the states in Display D2.8, you get the stemplot shown in Display D2.13. To better summarize this data, it is helpful if you rewrite the leaf values in ascending order after each stem. This creates a stemplot with *ordered leaves*. Notice that the stemplot includes a key at the top that helps the reader interpret the data values.

```
The decimal point is at the |

79 | 69
80 | 2
81 | 239
82 | 024
83 | 0256789
84 | 1
85 | 25899
86 | 25
87 | 24569
88 | 016679
89 | 56667
90 | 13334699
91 | 667
```

Display D2.13: Basic stemplot of high school completion rates.

A stemplot effectively groups the data into equal-sized intervals, or **bins**. For example, you see from the leaves of the first line of the stemplot that two high school completion rates are in the bin between 79.0 and 79.9 percent (specifically 79.6 and 79.9). Similarly, one rate falls in the bin between 80.0 and 80.9; three rates fall in the bin between 81.0 and 81.9; and so on. To emphasize the number of rates that fall into each bin, you can imagine replacing the leaves by bars where the length of each bar is proportional to the number of leaves. (See Display D2.14.)



Display D2.14: Stemplot of high school completion rates with leaves replaced by bars.

Note that, like the dotplot in Display D2.9, the stemplot still shows the shape of the distribution of high school completion rates. The overall shape is still skewed left, with the frequency of values decreasing as you move toward smaller values. (If you have trouble visualizing this, imagine turning the stemplot 90° counterclockwise so that the stems form a horizontal number line.) And you still see that the most of the values fall between 85 and 91 percent.

However, unlike a dotplot, a stemplot allows you to see the actual data values. For example, you see from the stemplot in Display D2.13 that the three largest high school completion rates are 91.6, 91.6, and 91.7 percent (corresponding to Alaska, Minnesota, and Wyoming). The dotplot in Display D2.9, in contrast, shows that the two highest rates are very close to 92 percent, but you would have to guess the decimal precision. So, one of the advantages of a stemplot is that you can see the shape and characteristics of the distribution without losing sight of the data values. For this reason, a stemplot is a useful and very detailed graph for relatively small data sets, say up to 50 values.

PRACTICE: CONSTRUCTING STEMPLOTS

When you construct a stemplot, a fundamental decision is how to divide each data value into the stem and leaf. Your decision may or may not lead to a “good” stemplot that makes it easy to see the data distribution.

The table below gives a variety of data about the states in the Midwest and Northeast sections of the U.S. Driver Rate is the number of licensed drivers per 1000 residents in August 2002; Farms is the number of farms (in thousands) in 2001; Density is the population per square mile of land area in 2000; and Elevation is the highest point in the state, measured as feet above sea level.

State	**Driver Rate (per 1000)**	**Farms (thousands)**	**Density (persons per square mile)**	**Elevation (feet above sea level)**
Illinois	641	76.0	224.6	1255
Indiana	654	63.0	170.5	1257
Iowa	667	93.5	52.3	1670
Kansas	710	63.0	52.9	4039
Michigan	697	52.0	175.9	1979
Minnesota	598	79.0	62.5	2301
Missouri	689	108.0	81.7	1772
North Dakota	715	30.3	9.2	3506
Ohio	723	78.0	277.8	1549
South Dakota	720	32.5	10.0	7242
Wisconsin	703	77.0	95.5	1951
Connecticut	779	3.9	706.9	2380
Maine	722	6.7	41.7	5267
Massachusetts	707	6.0	813.7	3487
New Hampshire	752	3.1	140.4	6288
New Jersey	672	9.6	1143.9	1803
New York	573	37.5	402.7	5344

Pennsylvania	670	59.0	274.2	3213
Rhode Island	624	0.7	1013.3	812
Vermont	831	6.6	66.3	4393

Display D2.15: Driver rates, number of farms, population density, and highest elevation for Northeast and Midwest states. (Source: *The World Almanac and Book of Facts, 2004*.)

1. Construct two stemplots for Driver Rate, as described below. Provide a key for each stemplot.

a. One stemplot where the stems are the hundreds and tens places and the leaves are the ones places. (For example, Illinois' rate of 641 would have a stem of 64 and a leaf of 1).

b. A second stemplot where the stems are the hundreds places and the leaves are the tens. (Illinois would have a stem of 6 and a leaf of 4.)

SPECIAL NOTE: It is common practice to record only a single digit for each leaf. Hence, for the stemplot in 1b, you drop the units places. Although you can either truncate the superfluous digits or round, it is usually quicker to truncate. Truncating also makes it easier to find a particular data value within the stemplot.

2. Which stemplot, 1a or 1b, gives a better display of the distribution of the Driver Rate data? Explain.

3. For each of the remaining variables (Farms, Density, and Elevation), construct a stemplot using the "best" division between stem and leaf. If the choice is not obvious, construct two stemplots using different divisions and choose the better graph.

HISTOGRAM

The **histogram** is a traditional method of graphing quantitative data, especially useful when you have a dataset with a large number of observational units. Similar to the modified stemplot in Display D2.15, a histogram uses bars to show the frequency of data values that fall into classes. To construct a histogram, you first create equal-sized bins that cover all of the data values; you classify all of the data values into these bins; and then you graph the bin frequencies using connected bars. One sticky point is what to do when a data value falls on the boundary between two bins; in this case, it is customary to place the data value into the bin on the right.

Recall that all of the high school completion rates in Display D2.8 fall between 79.6 and 91.7 percent. So, you could create these bins of width 2 starting from 78:

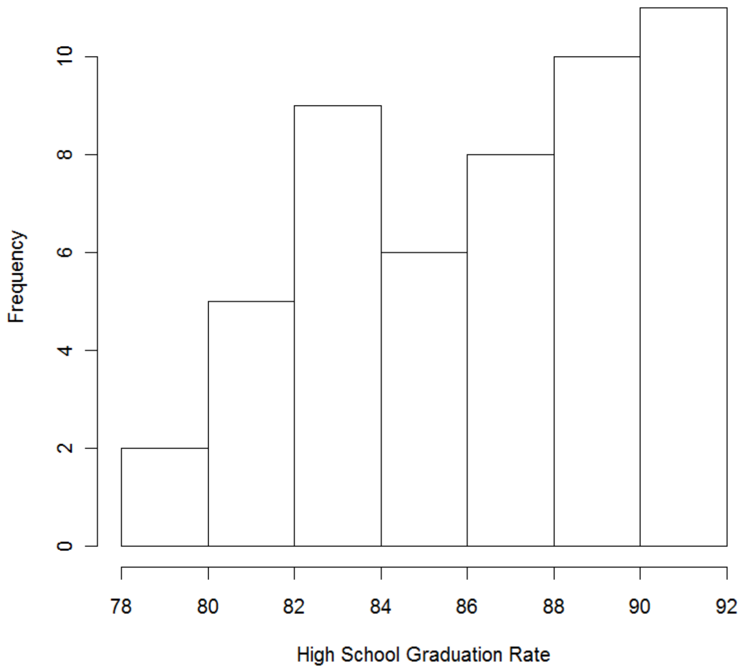
[78, 80), [80, 82), [82, 84), [84, 86), [86, 88), [88, 90), [90, 92)

Counting the number of completion rates that fall in [78, 80), [80, 82), and so on you obtain this frequency table.

Bin	[78,80)	[80,82)	[82,84)	[84,86)	[86,88)	[88,90)	[90,92)
Frequency	2	5	9	6	8	10	11

Display D2.16: Frequency table for high school completion rates.

This table tells us that two states have high school completion rates between 78 and 80 percent, five states have rates between 80 and 82 percent, and so on. To construct the histogram, you graph the frequencies (2, 5, 9, ...) against the bin boundaries using a bar chart in Display D2.17. Notice, however, that the bars touch each other, thereby implying that the data is a continuous quantitative variable rather than a discrete categorical variable.



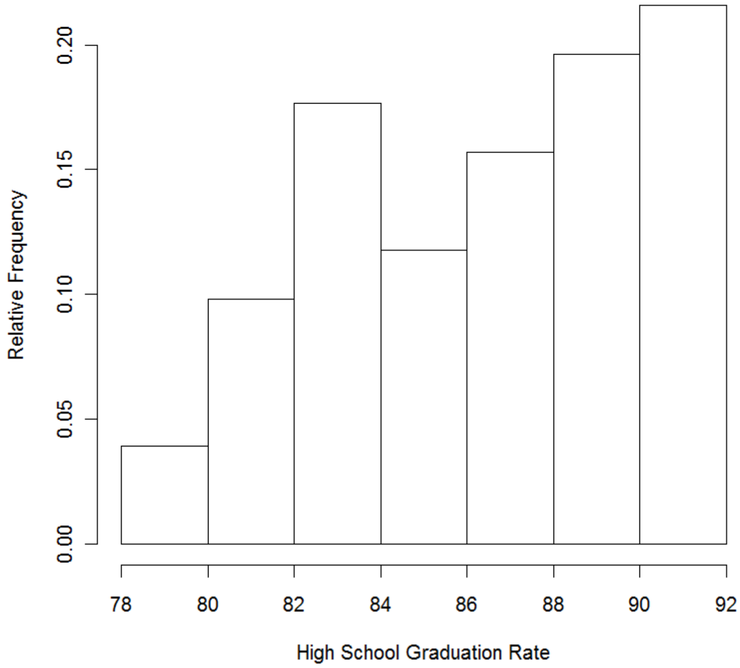
Display D2.17: Histogram of high school completion rates.

Sometimes it is helpful to further convert the bin frequencies to proportions, or **relative frequencies,** by dividing each frequency by the total number of values (51). Doing this for each bin, you get this **relative frequency table**:

Bin	[78,80)	[80,82)	[82,84)	[84,86)	[86,88)	[88,90)	[90,92)
Frequency	2	5	9	6	8	10	11
Proportion	0.039	0.098	0.176	0.118	0.157	0.196	0.216

Display D2.18: Relative frequency table of high school completion rates.

To construct a **relative frequency histogram,** you graph the proportions (0.039, 0.098, …) against the bin boundaries. Notice that the relative frequency histogram in Display D2.19 obeys the area principle; in fact, it looks identical to the histogram in Display D2.17 except for a change in the vertical axis labels. Also notice that the sum of all of the bars' heights is 1.



Display D2.19: Relative frequency histogram of high school completion rates.

These histograms provide a third view of the distribution of high school completion rates. Similar to what you previously saw from both the dotplot and stemplot, the histogram shows that the shape of the data is skewed left and most of the rates fall between 82 and 90 percent. One disadvantage of the histogram is that you lose sight of the actual data values within the bins. For example, the histogram in Display D2.17 shows that there are eleven states that have high school completion rates between 90 and 92 percent, but you don't know the values in this particular bin—all eight data values could be exactly the same or wildly different. In contrast, the dotplot in Display D2.9 at least shows you the distribution of the values within each interval; and the stemplot in Display D2.13 shows you that the actual values are 90.1, 90.3, 90.3, 90.3, 90.4, 90.6, 90.9, 90.9, 91.6, 91.6, and 91.7.   Nonetheless, the histogram remains a popular type of graph because it is a suitable and compact way to show the distribution of a large amount of quantitative data.

PRACTICE: INTERPRETING HISTOGRAMS

Look again at the relative frequency table for the high school completion rates with bins of width 2:

Bin	[78,80)	[80,82)	[82,84)	[84,86)	[86,88)	[88,90)	[90,92)
Frequency	2	5	9	6	8	10	11
Proportion	0.039	0.098	0.176	0.118	0.157	0.196	0.216

Display D2.20: Relative frequency table of high school completion rates.

1. What proportion of high school completion rates are less than 82 percent?

2. What proportion of rates is between 84 and 90 percent?

3. Suppose that you decide instead to use bins of width 4 starting at 78: [78, 82), [82, 86), [86, 90), [90, 94). Find the bin frequencies and bin proportions and complete this relative frequency table.

Bin	[78, 82)	[82, 86)	[86, 90)	[90, 94)
Frequency				
Proportion				

4. Construct a relative frequency histogram using bins in question 3.

5. Compare your four-bin relative frequency histogram in question 4 with the seven-bin relative frequency histogram in Display D2.20. Which is a "better" graph of the data? Explain.

6. REFLECTION You have seen that you can graph the high school completion rates using a dotplot, a stemplot, and a histogram.

(a) Which would you prefer if you had to make the graph by hand?

(b) If you could use a software package to make the graph, which would you prefer?

(c) If instead of exploring the high school completion rates of 51 rates you needed to explore the rates for the 254 counties in Texas, which graph might be better?

EXPERIMENT WITH DIFFERENT GRAPHS

When you construct a stemplot or a histogram, the data are grouped into bins, but you have to make choices about how that grouping is done. The choice of bin size can have a dramatic effect on the view of the data distribution, and some sizes are better than others for seeing the features of the distribution. It is good practice to experiment with several sizes for bins, choosing the one that seems to best represent the data.

To illustrate the use of different class sizes for stemplots, consider this table that gives the average price per gallon of regular gasoline for each U.S. state and the District of Columbia in June 2011.

State	Avg. Gas Price	State	Avg. Gas Price	State	Avg. Gas Price
Alaska	4.24	Kentucky	3.62	New York	3.95
Alabama	3.5	Louisiana	3.56	Ohio	3.61
Arkansas	3.52	Massachusetts	3.77	Oklahoma	3.57
Arizona	3.58	Maryland	3.71	Oregon	3.84
California	3.93	Maine	3.74	Pennsylvania	3.71
Colorado	3.66	Michigan	3.85	Rhode Island	3.84
Connecticut	4	Minnesota	3.64	South Carolina	3.43
District of Columbia	3.98	Missouri	3.5	South Dakota	3.75
Delaware	3.67	Mississippi	3.49	Tennessee	3.49

Florida	3.63	Montana	3.77	Texas	3.57
Georgia	3.59	North Carolina	3.6	Utah	3.61
Hawaii	4.04	North Dakota	3.78	Virginia	3.57
Iowa	3.66	Nebraska	3.75	Vermont	3.83
Idaho	3.71	New Hampshire	3.74	Washington	3.88
Illinois	3.99	New Jersey	3.68	Wisconsin	3.77
Indiana	3.74	New Mexico	3.61	West Virginia	3.75
Kansas	3.62	Nevada	3.71	Wyoming	3.66

Display D2.21: Average gasoline prices, June 2011. (Source: http://www.fuelgaugereport.com)

Suppose you first construct a stemplot by dividing each price so that the units and tenths digits form the stem and the hundredths digit form the leaf; the thousandths digit is truncated. For example, Kansas' price of $3.62 becomes 36 | 2. In general, this stemplot uses classes of width 0.10—that is, any value in the interval [3.70, 3.80) would be placed on the 37 stem. Then you get this stemplot of average gasoline prices:

```
1 | 2: represents 0.12
 leaf unit: 0.01

    3    34 | 399
   12    35 | 002677789
   25    36 | 0111223466678
  (14)   37 | 11114445557778
   12    38 | 34458
    7    39 | 3589
    3    40 | 04
         41 |
         42 | 4
```

Display D2.22: Stemplot of average gasoline prices.

From this graph, you see that the gasoline prices are slightly skewed right with most of the prices clustered in the $3.60's and $3.70's. But because 50 of the 51 data values are placed into only seven classes, you might be missing some features of the distribution that are hidden by so few classes. So, it may be worthwhile to try grouping the data using more classes.

In the stemplot in Display D2.22, all ten possible leaves (0, 1, 2, …, 9) were placed after a single stem. You can stretch the stemplot out by placing half of the leaves (0, 1, 2, 3, 4) after one stem, and the other half (5, 6, 7, 8, 9) after a duplicate stem. This is called a “5 leaves per stem” stemplot because there are five possible leaves after each stem. Using 5 leaves per stem changes the classes to width 0.05—that is, any value in the interval [3.40, 3.45) is placed on the first 34 stem, and any value in the interval [3.45, 3.50) is placed on the second 34 stem. If you use this alternative approach for these data, you get this new stemplot:

```
1 | 2: represents 0.12
 leaf unit: 0.01

 34 | 3
 34 | 99
 35 | 002
 35 | 677789
 36 | 01112234
 36 | 66678
 37 | 1111444
 37 | 5557778
 38 | 344
 38 | 58
 39 | 3
 39 | 589
 40 | 04
 41 |
 41 |
 42 | 4
```

Display D2.23: Stemplot of average gasoline prices with 5 leaves per stem.

In Display D2.23 you see more structure in these prices than you saw in Display D2.22. You see that more prices fell in the interval $3.60–$3.65 than in any other interval; a majority of the prices (27 out of 51, or 53%) fall between $3.60 and $3.80; there appears to be more detail and a smoother flow in the right-skewed shape; and you clearly see that the single highest price was $4.24 (corresponding to Alaska).

You can stretch the stemplot even further by placing the possible leaves after five stems: 0–1, 2–3, 4–5, 6–7, and 8–9. This stemplot variation is called “2 leaves per stem” (because there are two possible leaves after each stem) and changes the classes to width 0.02. Applying this approach, you obtain the following stemplot. (To shorten the display, we have placed Alaska’s gas price on a separate line labeled “HI”.)

```
 34 | 3
 34 |
 34 |
 34 | 99
 35 | 00
 35 | 2
 35 |
 35 | 6777
 35 | 89
 36 | 0111
 36 | 223
 36 | 4
 36 | 6667
 36 | 8
 37 | 1111
 37 |
 37 | 444555
 37 | 777
 37 | 8
 38 |
 38 | 3
 38 | 445
 38 |
 38 | 8
 39 |
 39 | 3
 39 | 5
 39 |
 39 | 89
 40 | 0
 40 |
 40 | 5

HI: 4.24
```

Display D2.24: Stemplot of average gasoline prices with 2 leaves per stem.

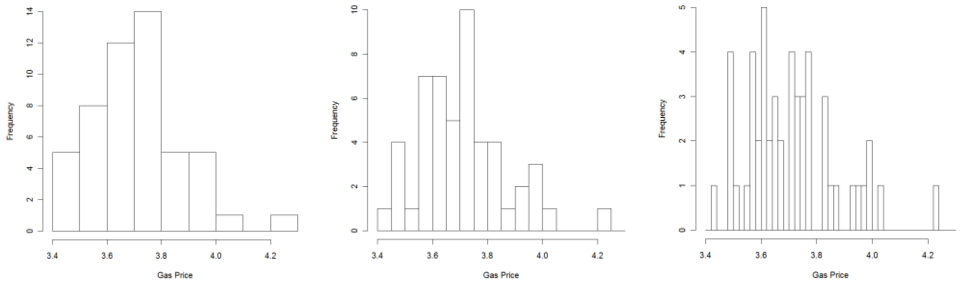
I think you'll agree that the stemplot in Display D2.24 is stretched too far—it is harder to see the basic right-skewed shape and there are now several gaps in the stemplot.

What is the best stemplot for these data? To answer this question, think about how well the each stemplot illustrates shape, center, spread, and interesting features of the distribution. Considering these elements, the stemplot in Display D2.23 with 5 leaves per stem appears to be the best—you see the basic shape of the distribution; you can make some meaningful observations about the center and spread; and you can see a few extreme prices that are much higher than the rest.

SPECIAL NOTE: The previous examples illustrate stemplots with 10, 5, and 2 leaves per stem. Could you construct a stemplot with 3 or 4 leaves per stem? No, because you want to have the same possible number of leaves after each stem. (Conceptually, this is the same idea as creating bins that each has the same width in a histogram.) In order for the ten possible leaf digits (0 through 9) to divide evenly among duplicate stems, you have to use a factor of ten. Because 10, 5, 2, and 1 are the only factors of ten, 3 or 4 leaves per stem will not work. The factors of ten also explain why 10 leaves per stem results in one of each stem, 5 leaves per stem results in two duplicates of each stem, and 2 leaves per stem results in five duplicates for each stem.

Similar to choosing class sizes for a stemplot, the choice of bin sizes is important when you construct a histogram. Although a computer program like Fathom will automatically select the bins according to a rule, it is helpful to experiment with wider or narrower bins to get a better graph of the data.

Here are three relative frequency histograms that use the same intervals for bins as were used for classes in the three stemplots above.



Display D2.25: Histograms of average gasoline prices using bins of width 0.10, 0.05, and 0.02.

The middle histogram, which uses bins of width 0.05, seems to best represent the distribution of gasoline prices of the 51 states. The first histogram on the left uses too few bins and you lose information about the distribution of gasoline prices within individual bins. The last histogram on the right, in contrast, uses too many bins and it is harder to see the basic shape of the distribution.

TECHNOLOGY LAB: CHOOSING THE BINS OF A HISTOGRAM

Fathom provides an easy way to modify the choice of bins in a histogram. By experimenting with the choice of bin width, you will understand the problems in constructing a histogram with bin widths that are either too small or too large.

PART A: Heights of college women

The datafile female_heights.txt contains the heights (recorded in inches) for 428 women taking introductory statistics at a Midwestern college one semester.

1. Import this data into Fathom.

2. Construct a histogram of the heights. Change the vertical scale to “density” by selecting the graph and choosing menu item Graph→Scale→Density.

(Note: By choosing the vertical scale of a histogram to be "density", the proportion of data values in a particular bin will be equal to the area of the corresponding bar of the histogram.)

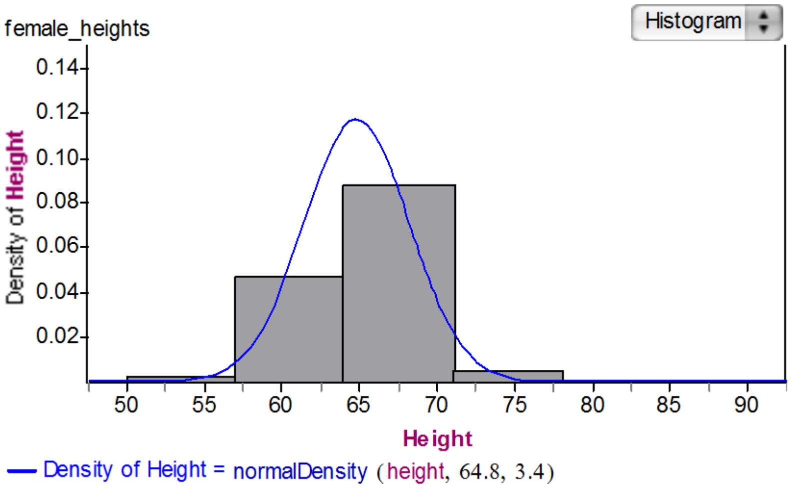
3. Note that the shape of these heights is symmetric. In a later topic, you will learn that the shape of this dataset can be well-described by a special bell-shaped curve called the normal density. With the histogram still selected, draw this normal curve on top of the histogram by selecting menu item Graph→Plot Function and typing `normaldensity(height,64.8,3.4)` in the formula editor window that appears.

Fathom give you two ways of changing the bins in a histogram:

- When you move the selection arrow between the bars of the histogram it changes to a double-headed arrow. When this happens, you can change the width of the bins by holding the mouse button and dragging. If you hold-and-drag on the left-most edge of the first bar, you can also change the starting value for the bins.
- Or, if you double-click inside the graph, a graph inspector window appears. Under the Properties tab, you can set the bin width by typing a new value for binWidth. You can also set the starting value for the bins by editing binAlignment. (Note: If your selections for binWidth and binAlignment result in the first few bins being empty, then Fathom will default to another value for binAlignment.)

The smooth curve represents the shape of the distribution of data if it contained infinitely many heights. The goal of this activity is to construct a histogram for the heights that is a good match for the smooth curve.

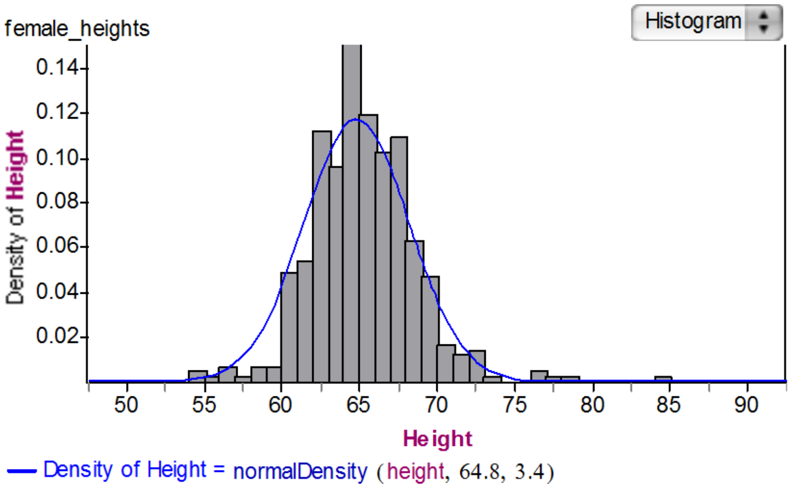
4. Using Fathom, choose only five bins of width 7 starting from 50; you should obtain a histogram such as the one pictured below.



Display D2.26: Fathom histogram of women heights with bin width 7 inches.

Notice that the histogram is not a good match to the curve. The curve gradually increases until a height of 65 inches and then gradually decreases for large values. In contrast, the histogram is not smooth as it takes big jumps at heights of 57 and 64 inches, and a big drop for the height of 71 inches. To make the histogram smoother, you need to choose a bin width smaller than 7 inches.

5. Redraw the histogram on Fathom using a small bin width of 1 inch. Your histogram should look similar to the histogram in Display D2.28 that uses approximately 30 bins of width 1 starting from 54 inches. The histogram appears to be a better match to the curve, but there is a new problem. We don’t see the big jumps and drops that we saw in the first histogram. But because the number of scores in each bin is small, the bar heights appear to be more erratic and the heights of the bars don’t follow the increasing and decreasing behavior of the curve.



Display D2.27: Fathom histogram of women heights with bin width of 1 inch.

To summarize the last two parts, your objective is to construct a histogram that

- is **a good match** to the underlying distribution curve (that is, the histogram doesn't have big jumps or big drops)
- is **relatively smooth** so you don't see much random fluctuation in the heights of the bars

6. Experiment with different bin widths and different starting values to find the settings that you think are "best." (*Hint:* Your bin width should be between 1 and 7 inches.)

7. Does the choice of the bin width and starting value depend on the number of data values in the dataset? To explore this question, the datafile female_heights_50.txt contains a smaller set of 50 women heights. Import these data into Fathom, create a histogram, and draw a bell-shape curve. Now find the "best" choices for bin width and starting value.

8. Based on your work, if you have more data, should you choose fewer (wider) bins or more (narrower) bins? Explain.

PART B: Histogram for skewed data.

Students in an introductory statistics class were asked the question: "How many movie dvds do you own." The datafile dvds.txt contains the response to this question for 636 students. These data can be modeled by a special right-skewed curve, an exponential density curve that you will learn more about in another topic.

1. Import these data into Fathom and construct a histogram. As before, select a density scale for the histogram.

2. Draw this distribution shape on top of the histogram by selecting menu item Graph→Plot Function and typing exponentialDensity(dvds, 27.3) in the formula editor.

3. Using the two criteria described above ("good match" and "smooth"), find a good choice for the bin width and starting value of the histogram.

PART C: Histogram for Old Faithful.

Old Faithful is a famous geyser in Yellowstone National Park, Wyoming. The waiting times between successive eruptions of Old Faithful follow a predictable pattern. The datafile oldfaithful.txt contains 106 times between eruptions.

1. Import these data into Fathom and construct a histogram. As before, when comparing histograms with different bin widths, it is helpful to choose the density scale.

2. Experiment with these bin widths and decide on the best choice. (Remember that on Fathom one can use exact values of the bin width using the Graph Inspector.)

10 9 8 7 6 5 4 3 2 1

3. Using your best choice of histogram, draw a smooth curve over the bar heights. Describe the shape of this smooth curve.

PRACTICE: EXPERIMENTING WITH DIFFERENT GRAPHS

Consider again the data on four variables for the states in the Midwest and Northeast regions of the U.S. (Display D2.16 on page XXX).

1. Construct three stemplots for Driver Rate data, as described below. For each data value, use the hundreds place as the stem and the tens place as the leaf. Don't forget to provide a key for each stemplot.

a. One stemplot that uses 10 leaves per stem. (This was the stemplot that was drawn in the previous practice activity.)

b. A second stemplot that uses 5 leaves per stem.

c. A third stemplot that uses 2 leaves per stem.

2. Which stemplot, 1a, 1b, or 1c, gives a better display of the distribution of the Driver Rate data? Explain.

3. Construct three histograms (or relative frequency histograms) for Elevation, as described below.

a. One histogram that uses bins of width 1000 starting from 0.

b. A second histogram that uses bins of width 500 starting from 0.

c. A third histogram that uses bins of width 250 starting from 0.

4. Which histogram, 3a, 3b, or 3c, gives a better display of the distribution of the Elevation data? Explain.

5. REFLECTION What does it mean for one graph to be a better display of the distribution of data than another? You can answer this question by drawing a "good" graph and a "bad" graph and explaining why the good graph is a better representation of the data than the bad graph.

ACTIVITY: THE SHAPE OF THE DATA

In this activity, you get some experience in collecting and graphing data, and studying the shapes of the corresponding distributions. You'll see data distributions that have a variety of shapes, including uniform, symmetric, and skewed.

Each person needs: a tennis ball, a single die, a centimeter ruler

1. Individually, devise a way to use a ruler to measure the diameter of a tennis ball in centimeters. Combine everyone's measurements into a class dataset.

2. Individually, count the number of rolls of a die until you observe all six possible outcomes (1, 2, 3, 4, 5, and 6) at least once. Combine everyone's results into a class dataset.

3. Individually, devise a way to use a ruler to measure the thickness of a single page of your textbook in centimeters. Combine everyone's measurements into a class dataset.

4. Refer to Display D2.28 on page XXX. Work individually or as a class to create a dataset of the first (left-most) digits of the counties' populations.

5. Use Display D2.28 to create dataset of the last (right-most) digits of the counties' populations.

Complete steps 6–9 for *each* of the five datasets that you collected above.

6. Construct a graph of the data.

7. Based on the graph, describe the shape of the distribution (uniform, symmetric, skewed right, skewed left, or other). Thinking about the type of data that was collected and the way it was collected, explain why the shape of the distribution “makes sense.”

8. Find the *mean* and the *median.* (These measures of center will be fully defined in Topic D3. Briefly, the mean is the arithmetic average calculated by adding all of the data values and dividing by the number of values. When the data values are put in chronological order, the median is the data value, or the mean of two data values, that is in the middle of the dataset.)

a. Is either measure of center significantly greater than the other, or are the mean and median approximately equal?

b. If the two measures of center are different, can you find any reasons why? Does it appear to be related to the shape of the distribution?

9. Describe any other interesting features (unusually large or small values, gaps, or clusters).

County	Population
Autauga County	43,671
Baldwin County	140,415
Barbour County	29,038
Bibb County	20,826
Blount County	51,024
Bullock County	11,714
Butler County	21,399
Calhoun County	112,249
Chambers County	36,583
Cherokee County	23,988
Chilton County	39,593
Choctaw County	15,922
Clarke County	27,867
Clay County	14,254
Cleburne County	14,123

County	Population
Graham County	33,489
Greenlee County	8,547
La Paz County	19,715
Maricopa County	3072,149
Mohave County	155,032
Navajo County	97,470
Pima County	843746
Pinal County	179,727
Santa Cruz County	38,381
Yavapai County	167,517
Yuma County	160,026
Arkansas County	20,749
Ashley County	24,209
Baxter County	38,386
Benton County	153,406

County	Population
Kern County	661,645
Kings County	129,461
Lake County	58,309
Lassen County	33,828
Los Angeles County	9,519,338
Madera County	123,109
Marin County	247,289
Mariposa County	17,130
Mendocino County	86,265
Merced County	210,554
Modoc County	9,449
Mono County	12,853
Monterey County	401,762
Napa County	124,279
Nevada County	92,033

Coffee County	43,615
Colbert County	54,984
Conecuh County	14,089
Coosa County	12,202
Covington County	37,631
Crenshaw County	13,665
Cullman County	77,483
Dale County	49,129
Dallas County	46,365
DeKalb County	64,452
Elmore County	65,874
Escambia County	38,440
Etowah County	103,459
Fayette County	18,495
Franklin County	31,223
Geneva County	25,764
Greene County	9,974
Hale County	17,185
Henry County	16,310
Houston County	88,787
Jackson County	53,926
Jefferson County	662,047
Lamar County	15,904
Lauderdale County	87,966
Lawrence County	34,803
Lee County	115,092
Limestone County	65,676
Lowndes County	13,473
Macon County	24,105
Madison County	276,700
Marengo County	22,539
Marion County	31,214
Marshall County	82,231
Mobile County	399,843
Monroe County	24,324
Montgomery	223,510

Boone County	33,948
Bradley County	12,600
Calhoun County	5,744
Carroll County	25,357
Chicot County	14,117
Clark County	23,546
Clay County	17,609
Cleburne County	24,046
Cleveland County	8,571
Columbia County	25,603
Conway County	20,336
Craighead County	82,148
Crawford County	53,247
Crittenden County	50,866
Cross County	19,526
Dallas County	9,210
Desha County	15,341
Drew County	18,723
Faulkner County	86,014
Franklin County	17,771
Fulton County	11,642
Garland County	88,068
Grant County	16,464
Greene County	37,331
Hempstead County	23,587
Hot Spring County	30,353
Howard County	14,300
Independence	34,233
Izard County	13,249
Jackson County	18,418
Jefferson County	84,278
Johnson County	22,781
Lafayette County	8,559
Lawrence County	17,774
Lee County	12,580
Lincoln County	14,492

Orange County	2,846,289
Placer County	248,399
Plumas County	20,824
Riverside County	1,545,387
Sacramento County	1,223,499
San Benito County	53,234
San Bernardino	1,709,434
San Diego County	2,813,833
San Francisco	776,733
San Joaquin County	563,598
San Luis Obispo	246,681
San Mateo County	707,161
Santa Barbara	399,347
Santa Clara County	1,682,585
Santa Cruz County	255,602
Shasta County	163,256
Sierra County	3,555
Siskiyou County	44,301
Solano County	394,542
Sonoma County	458,614
Stanislaus County	446,997
Sutter County	78,930
Tehama County	56,039
Trinity County	13,022
Tulare County	368,021
Tuolumne County	54,501
Ventura County	753,197
Yolo County	168,660
Yuba County	60,219
Adams County	363,857
Alamosa County	14,966
Arapahoe County	487,967
Archuleta County	9,898
Baca County	4,517
Bent County	5,998
Boulder County	291,288

County	
Perry County	11,861
Pickens County	20,949
Pike County	29,605
Logan County	22,486
Lonoke County	52,828
Madison County	14,243
Chaffee County	16,242
Cheyenne County	2,231
Clear Creek County	9,322

Display D2.28: Populations of a number of counties from Alabama, Alaska, Arizona, Arkansas, California, and Colorado from the 2000 U.S. Census. (Source: *http://www.census.gov/popest/counties/files/CO-EST2005-ALLDATA.csv*)

CLASSES OF DATA AND SHAPE

There are two general classes of quantitative data, called Counts/Amounts, and Measurements. The famous statistician John Tukey in his EDA (Exploratory Data Analysis) book gave special names to these classes since they tend to have common shapes. Some familiarity with these classes will be helpful in our exploration and summarization of data.

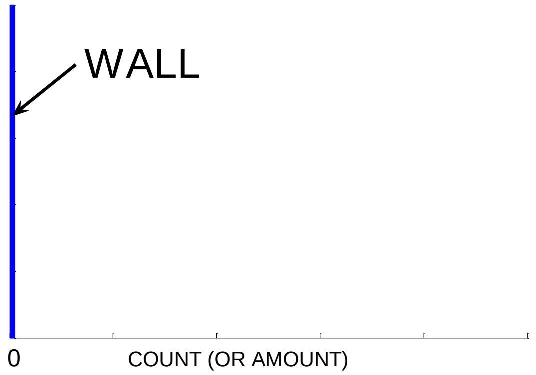
- COUNTS or AMOUNTS. You can collect the **count** of something, like the **number** of people in your family, the **number** of cookies you ate today, or the **number** of students who attend your school. Also you may be interested in the **amount** of money you spent on books this semester, the **amount** of time you spend studying for this course, or the **amount** of money you have in your saving account. Batches of counts or amounts are typically right-skewed.
- MEASUREMENTS. Data may be a collection of **measurements**, such as the **heights** of individuals of a single gender, the **time** it takes you to drive from home to school for many trips, or the **highest recorded temperature** for a city for many days in a particular month. Measurements are typically symmetric.

Why are counts and amounts typically right-skewed? As an example, the author collected the duration (in minutes) of 58 phone calls for a local company from the monthly bill:

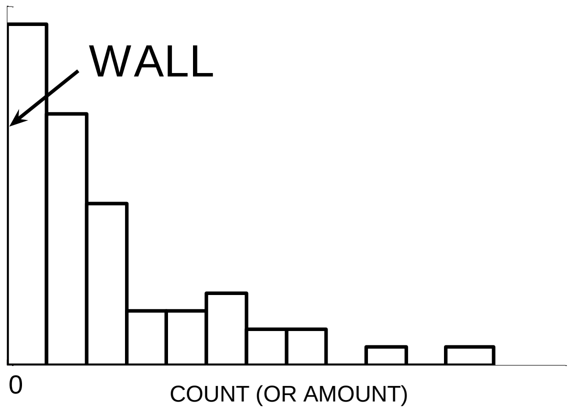
```
  2.8    1.8    0.9    0.3    1.2    1.2    2.4   12.2    1.5
```

```
5.3   0.9   0.6   4.3   1.0   0.7   0.5   1.2   1.2
6.9   0.6   0.3   4.5   0.6   7.8   0.8   0.9   0.8
5.3   7.5   4.2   2.7   1.5   2.4   6.0   3.6   6.3
1.2   2.7   1.1   0.6   0.4   2.7   1.0   1.4   2.1
0.6   3.1   1.6   1.5   3.0   5.7   0.3   9.9   1.7
3.7   0.6   2.0   3.0
```

Here the duration would be an example of an Amount – the amount of time that a phone call lasts. Note that a phone call's duration can't be negative, so all of the data must fall to the right of a "wall" of zero minutes. Also in many situations small positive values of Amounts are likely. It follows that the Amounts will tend to crowd up against the wall, resulting in a right-skewed distribution shape.



Indeed, if we graph the phone durations using a histogram, we see the right-skewed shape.

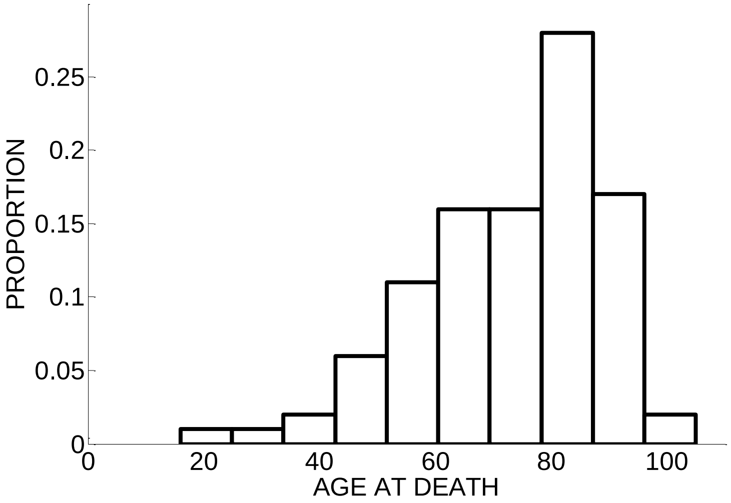


In contrast, consider a simple illustration of a batch of measurements. Suppose that each student in a class is asked to guess at an age of the instructor and suppose the instructor appears to be 35 years old. The error in a student's guess is defined to be

ERROR = GUESS – TRUE AGE = GUESS – 35.

Now some students will make guesses that are too small, resulting in negative errors, and other students will make guesses too high that result in positive errors. Most of the errors will be close to 0 and there will be (roughly) the same number of positive errors and negative errors. Here the distribution of measurements will be approximately symmetric. Unlike the first example, there is no wall that restricts the value of the error of the guess.

Is it possible for real-life data to be left-skewed? It is more likely to see right-skewed or symmetric data, but some variables have left-skewed distributions. This happens when data must fall to the left of a wall, where the wall represents the largest possible data value. One example of a left-skewed variable mentioned earlier would be test scores of an "easy" test. Here all scores must fall to the left of 100 percent and there will be many high grades and a trail of low grades. A second example would be age at death; a histogram of some ages of death collected from the obituary column of a local newspaper is displayed below. Note that most of the ages at death are between 70 and 90 years and the bins proportions decrease slowly for smaller values. This skewness is caused by a limit to the length of human life.



ACTIVITY: MATCHING VARIABLES AND SHAPES

DESCRIPTION: In this activity, you will get some experience matching histograms with different variables. Before doing this activity, it will be helpful to review the previous material on classes of data to understand some common distribution shapes.

For each group of variables, read the descriptions. Then write the variable name over the corresponding histogram.

GROUP 1 OF VARIABLES:

[HOME RUNS] The number of home runs hit for all baseball players who had at least 300 at-bats (opportunities) to hit. Home run numbers tend to be somewhat right-skewed – low home run counts are more common than large home run counts.

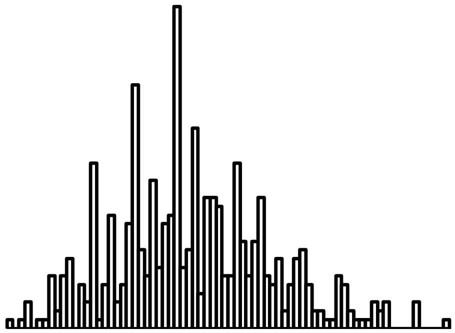
[FOOTBALL SCORES] The score of the winning team for a large number of (American) college football games. Particular football scores are popular, like 7, 14, etc.

[RUNNING TIMES] The times of women who ran in a marathon running race. Marathon running times are pretty symmetric, but it is more common to have a LARGE (slow) time than a SMALL (fast) time.

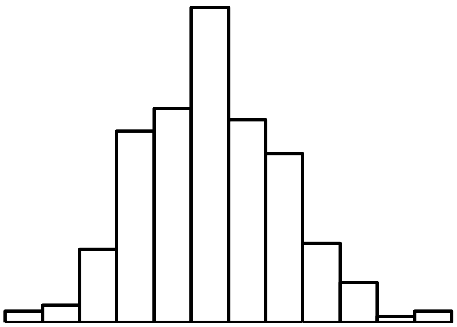
[BATTING AVERAGES] The batting averages for all baseball players who had at least 300 at-bats (opportunities) to hit. Batting averages tend to be very symmetric.

[SOCCER SCORES] The score of the winning team for a large number of soccer games played in club games in England. (If the game was tied, then the common score is recorded.) Soccer scores tend to be small.

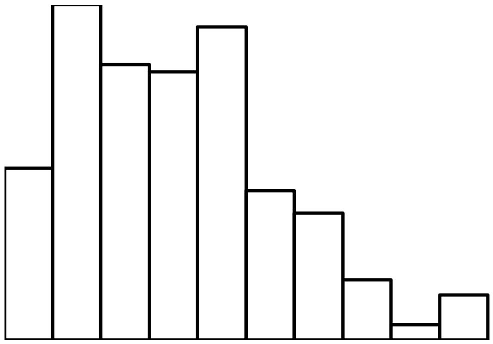
A



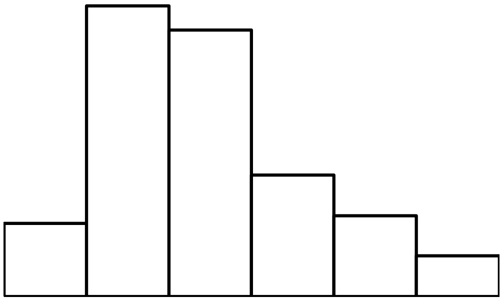
B



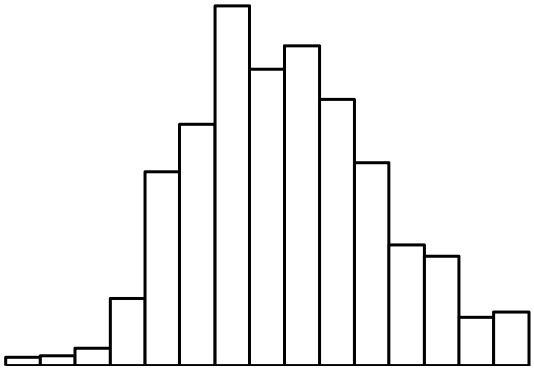
C



D



E



GROUP 2 OF VARIABLES:

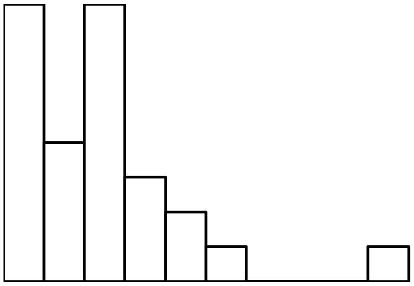
[DVDS] The number of movie DVDs owned by students in a college statistics class. Most students own few movie DVDs, but a couple of students own many DVDs.

[MILES FROM HOME] The miles from the school to the student’s hometown for students in a college statistics class. Students from this school tend to come from three regions of the state – one region within 30 miles of the school, a second region about 100 miles from the school, and a third region about 150 miles from the school.

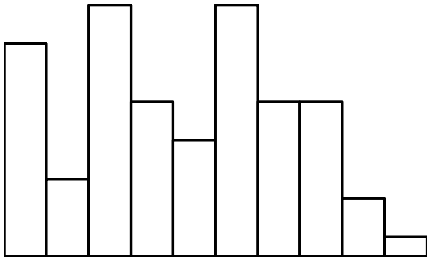
[TV] The average numbers of hours watching the television for students in a college statistics class. Most students watch only a little TV each day, but several students watch a lot of TV.

[SLEEP] The number of hours of sleep for students in a college statistics class. Students tend to get between 7 to 9 hours of sleep, but a few students got only a few hours of sleep.

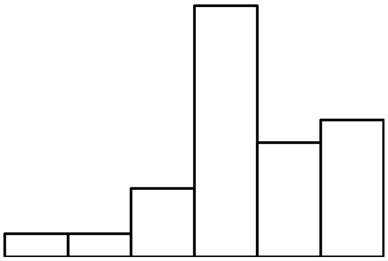
F



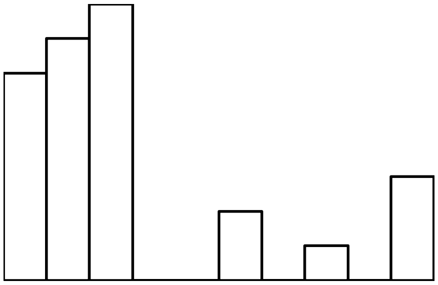
G



H



I



Classroom Capsule: Graphing Health and Liking School

Overview: The students will learn about different graphs for a single group of measurement data and understand principles for good graphs.

Objectives: The students will get experience drawing stemplots, dotplots, and histograms. They will understand that the construction of any graph involves choices, and the quality of the graphical display depends on these choices. For a given dataset, the students will learn that they can be several suitable graphs. The students will be able to criticize graphical displays presented in the media.

Description: In the UNICEF study, we focus on two variables that were measured for a group of 21 countries: “health”, the percentage of young people rating their health as "fair or poor", aged 11, 13, and 15, and “swell”, the percentage of young people 'liking school a lot', aged 11, 13, and 15. The table of the variables is shown below.

Country	health	swell
Austria	15.6	36.1
Belgium	13.1	17.9
Canada	13.7	21.9
Czech_Republic	11.8	11.6
Denmark	14.8	21.4
Finland	11.0	8.0
France	NA	21.7
Germany	14.9	29.5
Greece	10.1	29.5
Hungary	14.9	26.3
Ireland	12.9	22.3
Italy	12.5	13.0
Netherlands	17.2	34.4
Norway	18.5	38.9

Poland	14.4	17.3
Portugal	19.1	31.1
Spain	9.0	22.8
Sweden	13.2	21.6
Switzerland	9.1	22.3
United_Kingdom	22.6	19.0
United_States	19.8	23.4

1. For the health variable, construct three stemplots.

(a) First, construct a stemplot where you break between the units and tenths places and have ten leaves per stem

(b) Next, construct a stemplot where you break between the tens and units places and have two leaves per stem

(c) Last, construct a stemplot where you break between the tens and units places and have five leaves per stem

2. Among the three stemplots you drew in part 1, which is the best graph of the data? Why?

3. Using the best graph, write a descriptive paragraph of the data include statements about shape, center, spread and any unusual features.

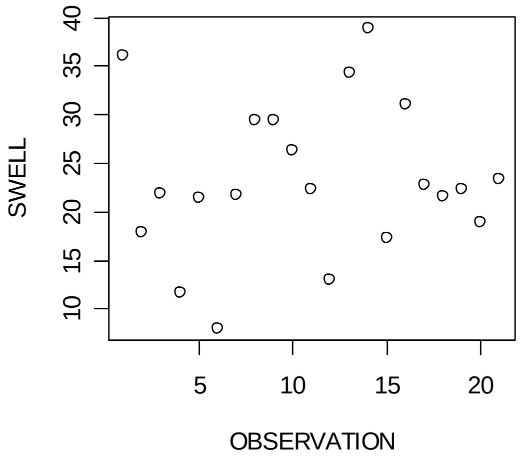
4. On the basis of the graph, how you would you describe American children's health in comparison with the 20 other countries in this study?

5. Four different graphs are drawn of the "swell" variable. The first graph is an index plot where the swell measurement is plotted as a function of the observation number and the remaining three graphs are histograms using different bin widths. Rank the four graphs in order (1, 2, 3, 4) from the "best" display and the "worst" display. In the spaces below give your rankings and explain why you gave these rankings.

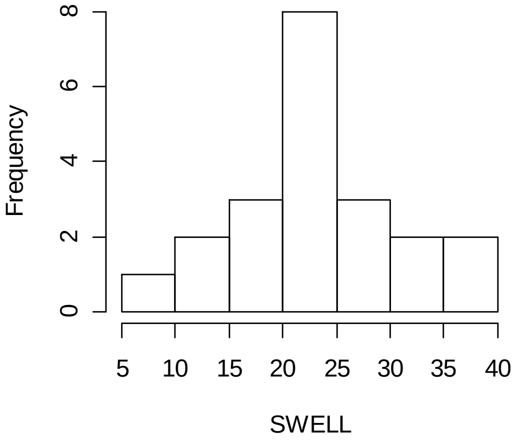
	GRAPH	WHY?
Best		
Next best		

Next best		
Worst		

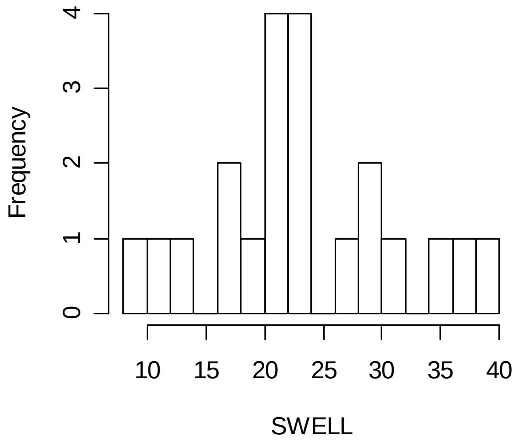
GRAPH A



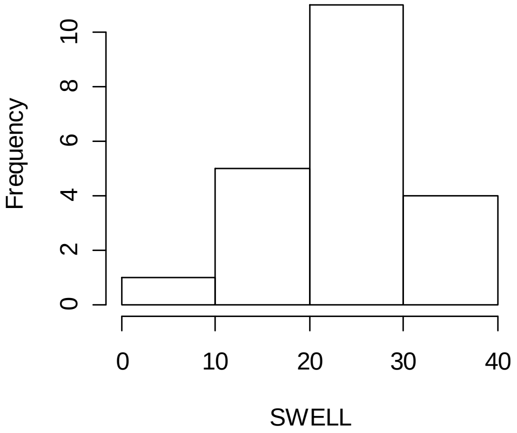
GRAPH B



GRAPH C



GRAPH D



Share and Summarize: One graphs data for a reason. Here the reason is to see the distribution of the data. If the data is graphed well, then it will be relatively easy to detect its shape, locate the center value, and make some statement about the spread of the values. It should be easy to summarize a dataset from looking at its graph.

Following are the three stemplots of the health variable. Which is the best graph?

1. Stemplot 1 has too many lines or bins. It is difficult to see the distributional shape and there are gaps in the display that may not be meaningful.
2. Stemplot 3 only uses four lines and it is difficult to see the distributional shape – over half of the values are found on a single line.
3. Stemplot 2 seems to be the best display among these three. The shape (roughly symmetric about 13%) is visible and there is one possible outlier (22%) at the high end.

```
 STEMPLOT 1          STEMPLOT 2          STEMPLOT 3


   9 | 01        0. | 99        0. | 99
  10 | 1         1* | 011       1* | 011223334444
  11 | 08         t | 22333     1. | 57899
  12 | 59         f | 44445     2* | 2
  13 | 127        s | 7
  14 | 4899      1. | 899       1|0 means 10%
  15 | 6         2* |
  16 |            t | 2
  17 | 2
  18 | 5            1|0 means 10%
  19 | 18
  20 |
  21 |
  22 | 6

9|0 means 9.0%
```

After the class has decided on the best graphical display, then talk about describing the health variable, including statements about shape, average value, and spread. The health variable for the United States is at the right end of the graph, indicating that the health of children in the U.S. is worse than most of the countries in the study.

In the discussion about part 5, Graph A that constructs an index plot of the observations, is not a good graph. In some cases, this might be useful, but the purpose of this graph is not obvious and it doesn't display the distribution of the data well. The three histograms (graphs B, C, D) use different choices for the number of bins. Histogram B is the best choice since it best shows the distributional shape and it is easiest to find the average value. A stemplot actually is a type of histogram where one sees the data values – choosing the right number of bins is the same issue as choosing the right stemplot.

Application or Extension: Use a graphing calculator to construct a histogram for a dataset of interest. First construct the "automatic" histogram where the bins are determined by the calculator. Then modify the histogram by choosing half as many bins, and by choosing twice as many bins. Confirm that the automatic choice for bins results in the best graphical display.

WRAP-UP

A first step in exploring a batch of data is to construct a suitable graph. For categorical data, **bar charts, segmented bar charts,** and **pie charts** are useful for comparing the **frequencies** of different categories. For quantitative data, **dotplots** and **stemplots** are easy-to-construct graphs that help you visualize the **distribution** of the data. **Histograms** are particularly helpful when you explore the distribution of a large volume of quantitative data. All of these graphs follow the **area principle** where the area of a object representing a data value is proportional to its frequency. In practice, it is useful to experiment with several graphs—such as stemplots with different numbers of **classes** or histograms with different numbers of **bins**—in order to find the graph that "best" represents the distribution.

When you look at the distribution of a batch of quantitative data, you are interested in its basic **shape,** a **center** value, the **spread** of values, and interesting features such as **outliers,** gaps, or clusters. Some basic distribution shapes include **uniform, symmetric, skewed right,** and **skewed left.** You will see in the next topic that the shape of a distribution influences the way that you summarize a dataset.

EXERCISES

1. **Religions of Countries**

Suppose you are taking a college class on world religions and you are asked about the predominant religions in the world. *The World Almanac* gives the chief religion for a large group of countries. To save time, you decide to record the chief religion for a sample of countries randomly selected from the almanac list with the hope that the distribution of major religions in this sample will be similar to the distribution of religions for the entire list of countries. (In the below list, Indigenous refers to a religion that is unique to that particular country.)

Country	**Chief Religion**
Afghanistan	Muslim
Antigue and Barbuda	Protestant
Azerbaijan	Muslim
Belarus	Eastern Orthodox
Bosnia and Herzegovina	Muslim
Burkina Faso	Muslim
Cape Verde	Roman Catholic
Columbia	Roman Catholic
Croatia	Roman Catholic
Denmark	Lutheran
Ecuador	Roman Catholic
Estonia	Lutheran
Gabon	Christian
Greece	Greek Orthodox
Guyana	Christian
India	Hindu
Israel	Jewish
Kenya	Protestant
Kyrgyzstan	Muslim
Libya	Muslim
Madagascar	Indigenous
Malta	Roman Catholic
Micronesia	Roman Catholic
Mozambique	Indigenous
Netherlands	Roman Catholic
Norway	Lutheran
Papua New Guinea	Indigenous
Portugal	Roman Catholic
Rwanda	Roman Catholic
San Marino	Roman Catholic
Seychelles	Roman Catholic
Somalia	Muslim
Suriname	Hindu
Taiwan	Buddhist
Trinidad and Tobago	Roman Catholic
Uganda	Roman Catholic
Uraguay	Roman Catholic
Vietnam	Buddhist

Display D2.29: Chief religions of 38 countries. (Source: *The World Almanac and Book of Facts 2005,* World Almanac Education Group, Inc., 2005)

a. What type of variable is Chief Religion?

b. Construct a frequency table for Chief Religion.

c. Suppose you decide to combine the religions into the categories Christianity-based (Christian, Protestant, Roman Catholic, Greek Orthodox, Lutheran, East-Orthodox), Muslim, and Other (all remaining religions). Construct a frequency table of these categories.

d. Construct a bar chart for Chief Religion using the categories of part c.

e. Based on your work, which chief religions are shared by the most countries in this group?

2. Types of High Schools in Fort Worth

The high schools in the United States can be categorized into three distinct types: **public** schools are administered and financed by the local and state government, **private** schools are not administered by the government, and **charter** schools are organized and controlled by educators, parents, or private groups with an expressed purpose or philosophy. Suppose a family is moving to Fort Worth, Texas and they are interested in the distribution of the three types of schools in this city. Display D2.30 gives the school type for the 66 high schools in Fort Worth.

public	private	public	public	public	public
private	private	public	charter	public	private
public	public	public	public	public	charter
public	private	charter	private	public	public
private	private	public	charter	public	private
public	private	private	private	public	public
public	private	private	private	public	private
charter	charter	private	private	private	private
public	public	public	private	public	charter
public	public	public	public	private	public
public	charter	charter	private	public	public

Display D2.30: School type for 66 high schools in Fort Worth, Texas. (Source: http://www.greatschools.net)

a. Construct a frequency table for School Type.

b. Construct a bar chart for School Type.

c. What is the most common school type among the high schools in Fort Worth?

d. Would it be accurate to say that over half of the Fort Worth high schools are public?

3. **Educational Attainment of Americans**

To get some insight about the education background of American adults, this table gives the highest level of education for a group of adult Americans, age 21 or older.

Level of Education	**Frequency**
4 or more years of college	12
1 to 3 years of college	20
Grade 12	17
Grade 11	2
Grade 10	4
Grade 9	1
Grade 5, 6, 7, or 8	5
None or preschool	2

Display D2.31: Educational attainment of 63 adults. (Source: Collected from the 2000 U.S. Census.)

a. Suppose you are interested in categorizing the Level of Education into the three groups: “not completed high school,” “completed high school,” and “some college.” Construct a frequency table using these three groups.

b. Construct a segmented bar chart for the frequencies you found in part 3a.

c. How many adults in this group had at least a high school education?

4. **Scores on the AP Calculus Exam**

Many high school students take the Advanced Placement (AP) Calculus course; this provides an opportunity for students to take a college-level calculus course while still in high school. At the end of the course, the student takes an AP Calculus Exam; if the student obtains a high grade on this exam, he or she can get credit for a college calculus course. There are two AP calculus exams that are given; the “AB exam” covers the content of the first calculus course in of and the “BC exam” covers the content of the first and second calculus courses. Display D2.32 gives the grade distribution of all students who took the AB exam and the BC exam in 2005.

AP Calculus AB Exam	
Exam Grade	Number of Examinees
Score of 5	38,539
Score of 4	36,347

AP Calculus BC Exam	
Exam Grade	Number of Examinees
Score of 5	23,877
Score of 4	9,237

Score of 3	33,006
Score of 2	31,141
Score of 1	46,959

Score of 3	10,929
Score of 2	3,695
Score of 1	6,677

Display D2.32: Grade distribution of all students who took the AB or BC forms of the AP calculus exam in 2005. (Source: The College Board, apcentral.collegeboard.com.)

a. Construct a histogram of the frequency of students obtaining the different grade levels for the AB exam.

b. What is the shape of this histogram?

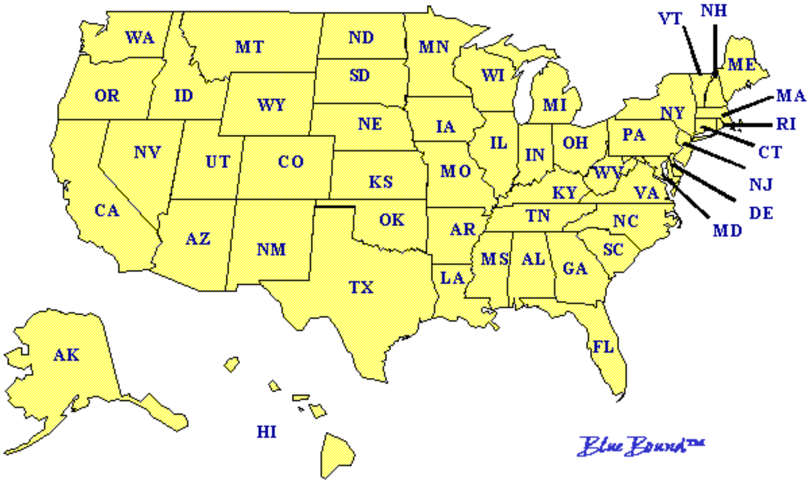
c. What is the most common score on the AB exam?

d. Answer parts a, b, and c for the grades on the BC exam.

e. Suppose that college credit is given if a student scores 3 or higher on the exam. From your work, estimate the proportion of students who would get college credit for each exam.

5. **How Many States Have You Visited?**

Look at the map below and count how many different U.S. states you have visited in your lifetime. Assume that "visited" means "stayed overnight."



a. What is the total number of states that you've visited?

Students in a statistics class were also asked, "How many different states have you visited in your lifetime?" Here are their answers:

46	10	15	14	8	10	13	11	18	13	13
12	10	11	5	30	32	11	25	14	12	11

Display D2.33: Number of states visited by 22 students.

b. Construct a dotplot of the data in Display D2.33.

c. Write a short paragraph that describes the distribution's shape, center, spread, and any interesting features.

d. Find an interval of values that approximately contains 50% of the data values.

e. Compare the number of states you have visited (from part 5a) with the distribution for the students in the statistics class. Can you consider yourself a "well-traveled person" relative to the students in this class? Explain.

6. **How Do You Sleep?**

Answer parts 6a, 6b, and 6c based on your own experience.

a. What time did you go to bed last night?

b. What time did you wake up this morning?

c. To the nearest quarter hour, how many hours of sleep did you get last night?

Students in a statistics class were also asked to calculate how many hours of sleep they each got last night. Here are the results:

7.00	6.50	10.00	6.25	9.50	8.50	6.00	8.00	8.75	5.50	9.00
9.38	5.75	5.50	8.25	8.50	8.00	8.50	7.80	8.75	7.75	6.00

Display D2.34: Hours of sleep for 22 students.

d. Construct a suitable graph of the data in Display D2.34.

e. Describe the distribution of this data, including shape, center, spread, and any interesting features.

f. How does your amount of sleep (from part 6c) compare with the distribution for the students in the statistics class? Would it be appropriate to say that you get an "average" amount of sleep? Explain.

Assume that a person's amount of sleep is categorized as "little" if it is less than 7 hours, "moderate" if it is greater than or equal to 7 hours but less than 9 hours, and "high" if it is greater than 9 hours.

g. Use the data in Display D2.34 and the categories "little," "moderate," and "high" to construct a frequency table and bar chart.

h. What proportion of students gets a "little" amount of sleep?

i. In this exercise you first treated the amount of sleep as a quantitative variable, and then as a categorical variable. Which treatment did you prefer? Why?

7. **Gross Sales of Movies Starring Julia Roberts**

The table below lists the U.S. gross sales of 29 movies featuring Julia Roberts.

Movie	**Gross Sales (millions of dollars)**
America's Sweethearts (2001)	94
Charlie Wilson's War (2007)	67
Closer (2004)	34
Conspiracy Theory (1997)	76
Duplicity (2009)	41
Dying Young (1991)	34
Eat Pray Love (2010)	81
Erin Brockovich (2000)	126
Flatliners (1990)	61
Full Frontal (2002)	3
Hook (1991)	120
I Love Trouble (1994)	31
Mary Reilly (1996)	6
Mexican, The (2001)	67
Michael Collins (1996)	11
Mona Lisa Smile (2003)	64
My Best Friend's Wedding (1997)	127
Notting Hill (1999)	116
Ocean's Eleven (2001)	183
Ocean's Twelve (2004)	126
Pelican Brief, The (1993)	101
Pretty Woman (1990)	178
Runaway Bride (1999)	152
Sleeping with the Enemy (1991)	102
Something to Talk About (1995)	51
Stepmom (1998)	91
Valentine's Day (2010)	111

Display D2.35: Gross sales of several Julia Roberts movies, as of June 2011. (Source: Internet Movie Database www.imdb.com.)

a. Construct a stemplot of the gross sales for these Julia Roberts movies by breaking between the tens and units digits, and using ten leaves per stem. (So for example, the first gross sales of 94 would have a stem of 9 and a leaf of 4. The gross sales of 6 or 06 would have a stem of 0 and a leaf of 6.)

b. Construct an alternative stemplot of the gross sales by again breaking between the tens and units digits and using five leaves per stem. Compare your two stemplots. Which one do you think is a better representation of the distribution of gross sales? Why?

c. Describe the features of the distribution of this data.

8. **How Random Are You?**

Students in a statistics class were asked to choose a number between 1 and 20. Here are the responses:

11	2	7	13	13	16	7	17	8	8	12
5	8	12	7	20	17	7	8	2	13	13

Display D2.36: Selection of numbers between 1 and 20 for 22 students.

a. Construct a stemplot of these numbers by breaking between the tens and units places and using five leaves per stem.

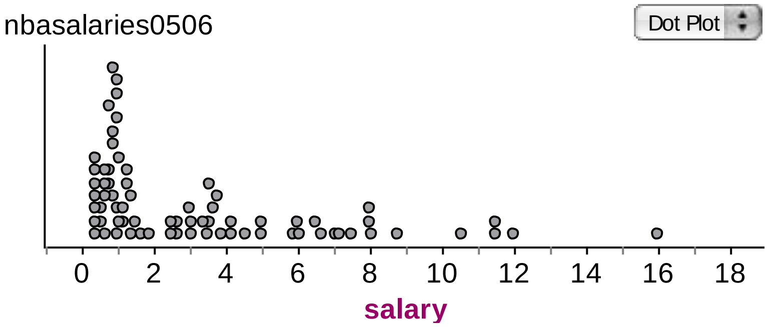
b. Can you find particular numbers that were popular among the students? Is there any explanation why these numbers might be popular?

c. Can you find numbers that were relatively unpopular? Is there any explanation why these numbers might be unpopular?

d. Do you think a distribution of random numbers (created by rolling a 20-sided die or by a computer algorithm) would look different from this distribution of student-chosen numbers? If so, explain how.

9. **Salaries of Basketball Players**

Only a small number of players have the opportunity to play for the NBA basketball league, but if they do get this opportunity, they receive very large salaries. The 2005-06 salaries (in millions of dollars) were collected for all players on five professional NBA teams (Hawks, Nuggets, Lakers, Hornets, Kings). There are a total of 72 players in this dataset. Display D2.37 shows a dotplot of these salaries.



Display D2.37: Dotplot of salaries (in millions of dollars) of players on five NBA teams for the 2005-06 season. (Source: USA Today, April 2006.)

a. Write a short paragraph describing the features of this distribution of salaries.

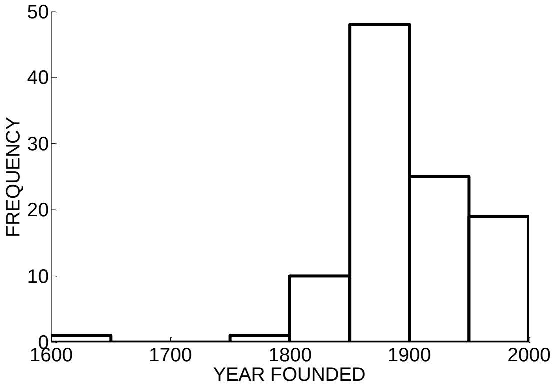
b. Estimate the number of players with a salary less than $2 million.

c. Estimate the most common salary among these players.

d. Suppose a "star" basketball player earns more than $10 million in this season. How many stars are there in this dataset? Given that this dataset includes the salaries for five teams, how many stars are there, on average, on each team?

10. **Year Founded for 104 Colleges**

It is interesting that Harvard University was founded in 1636, over one hundred years before the United States was founded. This motivates the question: "Generally, when were most of colleges in the United States founded?" To help answer this question, a histogram was constructed for the year founded for 104 U.S. colleges selected at random from a list of all colleges.



Display D2.38: Year founded for a selection of colleges. (Source: *U.S. News and World Report College*, 2004)

a. Describe the shape of these data.

b. Name a center value that represents a typical year founded for these schools.

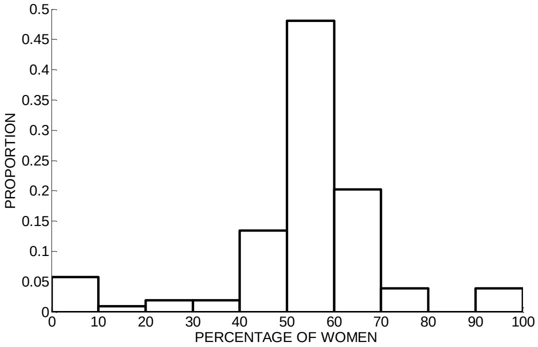
c. There is one outlier in the data corresponding to Harvard University that was founded in 1636. Can you guess at the locations of the colleges that were founded between 1750 and 1800?

d. Estimate the number of schools that were founded before 1900.

e. Estimate the number of schools that were founded in 1950 or later.

11. **Percentage of Women for 104 Colleges**

Many college freshmen are interested in the ratio of men to women at their school. This relative frequency histogram shows the percentage of women enrolled at the same 104 schools randomly selected for exercise 10:



Display D2.39: Percentage of women enrolled at a selection colleges. (Source: *U.S. News and World Report College*, 2004)

a. Describe the shape, center, spread, and any interesting features of this distribution.

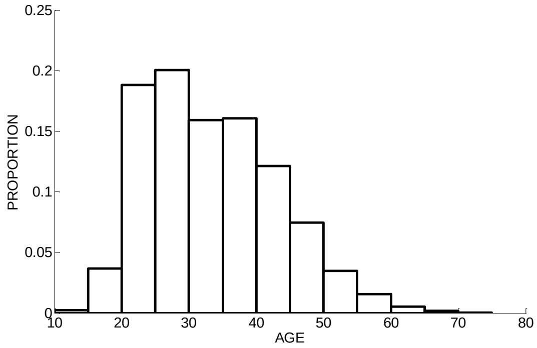
b. What proportion of schools has 50 percent or more women?

c. Based on this graph, do you believe that there are more women or men attending colleges in the United States? Explain.

d. What explanations can you provide for the colleges that have unusually small or large proportions of women?

12. **Ages of Women Participating in a Marathon**

The marathon race is unique in that it includes participants from a wide range of ages. This relative frequency histogram shows the ages of women participating in the Grandma's Marathon race in Duluth, Minnesota.



Display D2.40: Ages of women participating in the 2003 Grandma's Marathon, Duluth, Minnesota (Source: www.grandmamarathon.com)

a. Describe the shape of this distribution of ages.

b. What is a center value for the ages of women in this marathon?

c. What proportion of these women were age 20 or older, but less than age 30?

d. Find an interval that contains approximately the middle 50% of the runners' ages.

e. If a total of 10,000 women competed in this marathon, how many women were older than 50?

13. **Fares of Airplane Flights to Different Cities**

The National Council of Teachers of Mathematics is considering hosting a national conference in Detroit, Michigan. Assuming that participants may come from all of the United States, what is the fare for a round-trip airplane flight to the conference? To explore this question, consider this table of fares from Detroit, Michigan.

City	Fare ($)
Boston, MA	327
Chicago, IL	92
Denver, CO	198
Fairbanks, AK	696
Fargo, ND	369
Honolulu, HI	701
Houston, TX	219
Kansas City, MO	220

City	Fare ($)
New Orleans, LA	280
New York, NY	170
Orlando, FL	236
Philadelphia, PA	258
Phoenix, AZ	204
Portland, OR	352
Raleigh, NC	224
San Diego, CA	312

Las Vegas, NV	242	San Francisco, CA	310
Miami, FL	252	Sante Fe, NM	521

Display D2.41: Lowest-cost round-trip airplane fares from Detroit, Michigan, April 2004. (Source: *www.orbitz.com*)

a. Construct a dotplot, stemplot, and histogram for these fares. Which graph do you think best represents the data, and why?

b. Describe the distribution of fares, including shape, center, spread, and any interesting features.

c. Why is there so much spread in plane fares? Give several reasons for this spread.

14. **Nutritional Content of Ben & Jerry's Ice Cream**

Ben & Jerry's is one of the most famous makers of gourmet ice cream. The table below gives nutritional information for some of their flavors:

Flavor	**Calories**	**Sodium (grams)**	**Flavor**	**Calories**	**Sodium (grams)**
Brownie Batter	310	115	Karamel Sutra®	280	75
Butter Pecan	290	80	Mint Chocolate Cookie	270	100
Cherry Garcia®	260	50	New York Super Fudge Chunk®	310	55
Chocolate	260	50	Oatmeal Cookie Chunk	280	120
Chocolate Chip Cookie Dough	270	90	One Sweet Whirled	280	85
Chocolate Fudge Brownie™	270	80	Peanut Butter Cup™	380	140
Chubby Hubby®	330	160	Phish Food®	280	90
Chunky Monkey®	300	45	Pistachio Pistachio®	280	125
Coffee	240	60	Primary Berry Graham	270	110
Coffee HEATH® Bar Crunch	290	115	Strawberry	240	50
Dublin Mudslide™	270	80	Uncanny Cashew™	290	130
Everything But The... ®	320	80	Vanilla	240	55
Fudge Central®	300	60	Vanilla HEATH® Bar Crunch	300	120
Half Baked®	280	90	Vanilla Swiss Almond	280	65

Display D2.42: Calories and sodium in half-cup servings of select Ben & Jerry's ice cream flavors. (Source: *www.benjerry.com*)

a. Construct a histogram of Calories.

b. Write a short paragraph about the distribution of Calories, including shape, center, spread, and any interesting features.

c. What is your favorite flavor? How does the amount of calories in your flavor compare to the distribution of calories for the flavors in Display D2.42? (Is it average, below average, or above average?) (*Note:* If your favorite flavor and/or brand is not listed, try using the Internet to find its nutritional information.)

d. Construct a suitable graph for sodium.

e. Describe the distribution of sodium.

f. How does the amount of sodium in your favorite flavor compare to the distribution?

g. What explanations can you provide for the flavors that have unusually high amounts of sodium?

15. Data Shapes

Suppose you collect each of the following variables for each student in a high school class. Describe the expected shape of the distribution of the variable.

a. The number of first cousins.

b. The amount of money spent for the last haircut.

c. The pulse rate (number of beats in one minute).

d. The number of homework study hours in a week.

e. The composite score on the ACT exam.

16. Data Shapes

For each of the following data variables, construct a hypothetical graph of the distribution and describe its shape.

a. The month of the birthday (a number between 1 and 12) is recorded for a group of 1000 people.

b. The salaries for 50 people who work in a local company is collected.

c. A test with 100 possible points is administered to a large group of students. A majority of the students score more than 90 points.

d. The number of points scored by a basketball team is recorded for a large number of games.