CHAPTER

1

Looking at Data—Distributions

Students planning a referendum on college fees. See Example 1.1.

Introduction

Statistics is the science of learning from data. Data are numerical facts. Here is an example of a situation where students used the results of a referendum to convince their university Board of Trustees to make a decision.

EXAMPLE

1.1 Students vote for service learning scholarships. According to the National Service-Learning Clearinghouse: "Service-learning is a teaching and learning strategy that integrates meaningful community service with instruction and reflection to enrich the learning experience, teach civic responsibility, and strengthen communities."[1] University of Illinois at Urbana–Champaign students decided that they wanted to become involved in this national movement. They proposed a $15.00 per semester Legacy of Service and Learning Scholarship fee. Each year, $10.00 would be invested in an endowment and $5.00 would be used to fund current-use scholarships. In a referendum, students voted 3785 to 2977 in favor of the proposal. On April 11, 2006, the university Board of Trustees approved the proposal. Approximately $370,000 in current-use scholarship funds will be generated each year, and with the endowment, it is expected that in 20 years there will be more than a million dollars per year for these scholarships.

To learn from data, we need more than just the numbers. The numbers in a medical study, for example, mean little without some knowledge of the goals of the study and of what blood pressure, heart rate, and other measurements contribute to those goals. That is, *data are numbers with a context,* and we need to understand the context if we are to make sense of the numbers. On the other hand, measurements from the study's several hundred subjects are of little value to even the most knowledgeable medical expert until the tools of statistics organize, display, and summarize them. We begin our study of statistics by mastering the art of examining data.

Variables

Any set of data contains information about some group of *individuals*. The information is organized in *variables*.

INDIVIDUALS AND VARIABLES

Individuals are the objects described in a set of data. Individuals are sometimes people. When the objects that we want to study are not people, we often call them **cases.**

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

EXAMPLE

1.2 Data for students in a statistics class. Figure 1.1 shows part of a data set for students enrolled in an introductory statistics class. Each row gives the data on one student. The values for the different variables are in the columns. This data set has eight variables. ID is an identifier for each student. Exam1, Exam2, Homework, Final, and Project give the points earned, out of a total of 100 possible, for each of these course requirements. Final grades are based on a possible 200 points for each exam and the final, 300 points for Homework, and 100 points for Project. TotalPoints is the variable that gives the composite score. It is computed by adding 2 times Exam1, Exam2, and Final, 3 times Homework plus 1 times Project. Grade is the grade earned in the course. This instructor used cut-offs of 900, 800, 700, etc. for the letter grades.

Microsoft Excel

	A	B	C	D	E	F	G	H
1	ID	Exam1	Exam2	Homework	Final	Project	TotalPoints	Grade
2	101	89	94	88	87	95	899	B
3	102	78	84	90	89	94	866	B
4	103	71	80	75	79	95	780	C
5	104	95	98	97	96	93	962	A
6	105	79	88	85	88	96	861	B

FIGURE 1.1 Spreadsheet for Example 1.2.

spreadsheet

The display in Figure 1.1 is from an Excel **spreadsheet.** Most statistical software packages use similar spreadsheets and many are able to import Excel spreadsheets.

USE YOUR KNOWLEDGE

1.1 Read the spreadsheet. Refer to Figure 1.1. Give the values of the variables Exam1, Exam2, and Final for the student with ID equal to 104.

1.2 Calculate the grade. A student whose data do not appear on the spreadsheet scored 88 on Exam1, 85 on Exam2, 77 for Homework, 90 on the Final, and 80 on the Project. Find TotalPoints for this student and give the grade earned.

Spreadsheets are very useful for doing the kind of simple computations that you did in Exercise 1.2. You can type in a formula and have the same computation performed for each row.

Note that the names we have chosen for the variables in our spreadsheet do not have spaces. For example, we could have used the name "Exam 1" for the first exam score rather than Exam1. In many statistical software packages, however, spaces are not allowed in variable names. For this reason, when creating spreadsheets for eventual use with statistical software, it is best to avoid spaces in variable names. Another convention is to use an underscore (_) where you would normally use a space. For our data set, we could use Exam_1, Exam_2, and Final_Exam.

In practice, any set of data is accompanied by background information that helps us understand the data. When you plan a statistical study or explore data from someone else's work, ask yourself the following questions:

1. **Why? What purpose** do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than those for whom we actually have data?
2. **Who?** What **individuals** do the data describe? **How many** individuals appear in the data?
3. **What?** How many **variables** do the data contain? What are the **exact definitions** of these variables? Some variables have units. Weights, for example, might be recorded in pounds, in thousands of pounds, or in kilograms. For these kinds of variables, you need to know the **unit of measurement.**

EXAMPLE

1.3 Individuals and variables. The data set in Figure 1.1 was constructed to keep track of the grades for students in an introductory statistics course. The individuals are the students in the class. There are 8 variables in this data set. These include an identifier for each student and scores for the various course requirements. There are no units for ID and grade. The other variables all have "points" as the unit.

Some variables, like gender and college major, simply place individuals into categories. Others, like height and grade point average, take numerical values

for which we can do arithmetic. It makes sense to give an average salary for a company's employees, but it does not make sense to give an "average" gender. We can, however, count the numbers of female and male employees and do arithmetic with these counts.

CATEGORICAL AND QUANTITATIVE VARIABLES

A **categorical variable** places an individual into one of two or more groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

The **distribution** of a variable tells us what values it takes and how often it takes these values.

EXAMPLE

1.4 Variables for students in a statistics course. Suppose the data for the students in the introductory statistics class were also to be used to study relationships between student characteristics and success in the course. For this purpose, we might want to use a data set like the spreadsheet in Figure 1.2. Here, we have decided to focus on the TotalPoints and Grade as the outcomes of interest. Other variables of interest have been included: Gender, PrevStat (whether or not the student has taken a statistics course previously), and Year (student classification as first, second, third, or fourth year). ID is a categorical variable, total points is a quantitative variable, and the remaining variables are all categorical.

Microsoft Excel

	A	B	C	D	E	F
1	ID	TotalPoints	Grade	Gender	PrevStat	Year
2	101	899	A	F	Yes	4
3	102	866	B	M	Yes	3
4	103	780	C	M	No	3
5	104	962	A	M	No	1
6	105	861	B	F	No	4

FIGURE 1.2 Spreadsheet for Example 1.4.

In our example, the possible values for the grade variable are A, B, C, D, and F. When computing grade point averages, many colleges and universities translate these letter grades into numbers using A = 4, B = 3, C = 2, D = 1, and F = 0. The transformed variable with numeric values is considered to be quantitative because we can average the numerical values across different courses to obtain a grade point average.

Sometimes, experts argue about numerical scales such as this. They ask whether or not the difference between an A and a B is the same as the difference between a D and an F. Similarly, many questionnaires ask people to

respond on a 1 to 5 scale with 1 representing strongly agree, 2 representing agree, etc. Again we could ask about whether or not the five possible values for this scale are equally spaced in some sense. From a practical point of view, the averages that can be computed when we convert categorical scales such as these to numerical values frequently provide a very useful way to summarize data.

USE YOUR KNOWLEDGE

1.3 Apartment rentals. A data set lists apartments available for students to rent. Information provided includes the monthly rent, whether or not cable is included free of charge, whether or not pets are allowed, the number of bedrooms, and the distance to the campus. Describe the individuals or cases in the data set, give the number of variables, and specify whether each variable is categorical or quantitative.

Measurement: know your variables

The context of data includes an understanding of the variables that are recorded. Often the variables in a statistical study are easy to understand: height in centimeters, study time in minutes, and so on. But each area of work also has its own special variables. A psychologist uses the Minnesota Multiphasic Personality Inventory (MMPI), and a physical fitness expert measures "VO2 max," the volume of oxygen consumed per minute while exercising at your maximum capacity. Both of these variables are measured with special **instruments.** VO2 max is measured by exercising while breathing into a mouthpiece connected to an apparatus that measures oxygen consumed. Scores on the MMPI are based on a long questionnaire, which is also an instrument. Part of mastering your field of work is learning what variables are important and how they are best measured. Because details of particular measurements usually require knowledge of the particular field of study, we will say little about them.

instrument

Be sure that each variable really does measure what you want it to. A poor choice of variables can lead to misleading conclusions. Often, for example, the **rate** at which something occurs is a more meaningful measure than a simple count of occurrences.

rate

EXAMPLE

1.5 Accidents for passenger cars and motorcycles. The government's Fatal Accident Reporting System says that 27,102 passenger cars were involved in fatal accidents in 2002. Only 3339 motorcycles had fatal accidents that year.[2] Does this mean that motorcycles are safer than cars? Not at all—there are many more cars than motorcycles, so we expect cars to have a higher *count* of fatal accidents.

A better measure of the dangers of driving is a *rate,* the number of fatal accidents divided by the number of vehicles on the road. In 2002, passenger cars had about 21 fatal accidents for each 100,000 vehicles registered. There were about 67 fatal accidents for each 100,000 motorcycles registered. The rate for motorcycles is more than three times the rate for cars. Motorcycles are, as we might guess, much more dangerous than cars.

1.1 Displaying Distributions with Graphs

exploratory data analysis

Statistical tools and ideas help us examine data in order to describe their main features. This examination is called **exploratory data analysis.** Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two basic strategies that help us organize our exploration of a set of data:

- Begin by examining each variable by itself. Then move on to study the relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will follow these principles in organizing our learning. This chapter presents methods for describing a single variable. We will study relationships among several variables in Chapter 2. Within each chapter, we will begin with graphical displays, then add numerical summaries for more complete description.

Graphs for categorical variables

The values of a categorical variable are labels for the categories, such as "female" and "male." The **distribution** of a categorical variable lists the categories and gives either the **count** or the **percent** of individuals who fall in each category. For example, how well educated are 30-something young adults? Here is the distribution of the highest level of education for people aged 25 to 34 years:[3]

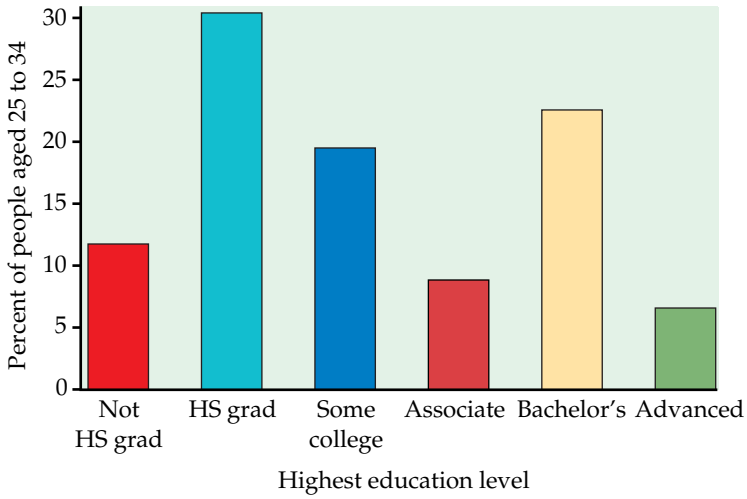
Education	Count (millions)	Percent
Less than high school	4.6	12.1
High school graduate	11.6	30.5
Some college	7.4	19.5
Associate degree	3.3	8.7
Bachelor's degree	8.6	22.6
Advanced degree	2.5	6.6

Are you surprised that only 29.2% of young adults have at least a bachelor's degree?

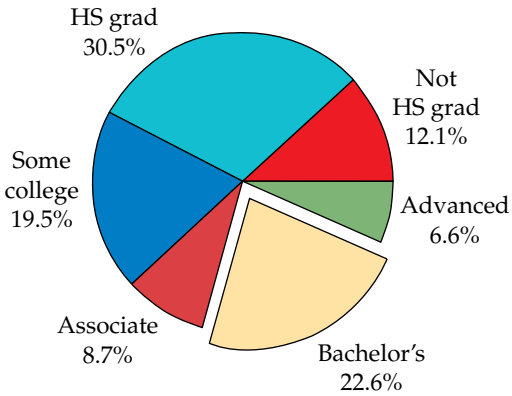
bar graph

pie chart

The graphs in Figure 1.3 display these data. The **bar graph** in Figure 1.3(a) quickly compares the sizes of the six education groups. The heights of the bars show the percents in the six categories. The **pie chart** in Figure 1.3(b) helps us see what part of the whole each group forms. For example, the "Bachelor's" slice makes up 22.6% of the pie because 22.6% of young adults have a bachelor's degree but no higher degree. We have moved that slice out to call attention to it. Because pie charts lack a scale, we have added the percents to the labels for the slices. *Pie charts require that you include all the categories that make up a whole. Use them only when you want to emphasize each category's relation to the whole.* Bar graphs are easier to read and are also more flexible. For example, you can use a bar graph to compare the numbers of students at your college majoring in biology, business, and political science. A pie chart cannot make this comparison because not all students fall into one of these three majors.



(a)



(b)

FIGURE 1.3 (a) Bar graph of the educational attainment of people aged 25 to 34 years. (b) Pie chart of the education data, with bachelor's degree holders emphasized.

USE YOUR KNOWLEDGE

1.4 Read the pie chart. Refer to Figure 1.3(b). What percent of young adults have either an associate degree or a bachelor's degree?

Bar graphs and pie charts help an audience grasp a distribution quickly. They are, however, of limited use for data analysis because it is easy to understand data on a single categorical variable, such as highest level of education, without a graph. We will move on to quantitative variables, where graphs are essential tools.

Data analysis in action: don't hang up on me

Many businesses operate call centers to serve customers who want to place an order or make an inquiry. Customers want their requests handled thoroughly. Businesses want to treat customers well, but they also want to avoid wasted time on the phone. They therefore monitor the length of calls and encourage their representatives to keep calls short. Here is an example of the difficulties this policy can cause.

EXAMPLE

1.6 Individuals and variables for the customer service center. We have data on the length of all 31,492 calls made to the customer service center of a small bank in a month. Table 1.1 displays the lengths of the first 80 calls. The file for the complete data set is *eg01-004,* which you can find on the text CD and Web site.[4]

Take a look at the data in Table 1.1. The numbers are meaningless without some background information. The *individuals* are calls made to the bank's call center. The *variable* recorded is the length of each call. The *units* are

TABLE 1.1

Service times (seconds) for calls to a customer service center

77	289	128	59	19	148	157	203
126	118	104	141	290	48	3	2
372	140	438	56	44	274	479	211
179	1	68	386	2631	90	30	57
89	116	225	700	40	73	75	51
148	9	115	19	76	138	178	76
67	102	35	80	143	951	106	55
4	54	137	367	277	201	52	9
700	182	73	199	325	75	103	64
121	11	9	88	1148	2	465	25

seconds. We see that the call lengths vary a great deal. The longest call lasted 2631 seconds, almost 44 minutes. More striking is that 8 of these 80 calls lasted less than 10 seconds. What's going on?

Figure 1.4 is a histogram of the lengths of all 31,492 calls. We did not plot the few lengths greater than 1200 seconds (20 minutes). As expected, the graph shows that most calls last between about a minute and 5 minutes, with some lasting much longer when customers have complicated problems. More striking is the fact that 7.6% of all calls are no more than 10 seconds long. It turned out that the bank penalized representatives whose average call length was too long—so some representatives just hung up on customers in order to bring their average length down. Neither the customers nor the bank were happy about this. The bank changed its policy, and later data showed that calls under 10 seconds had almost disappeared.

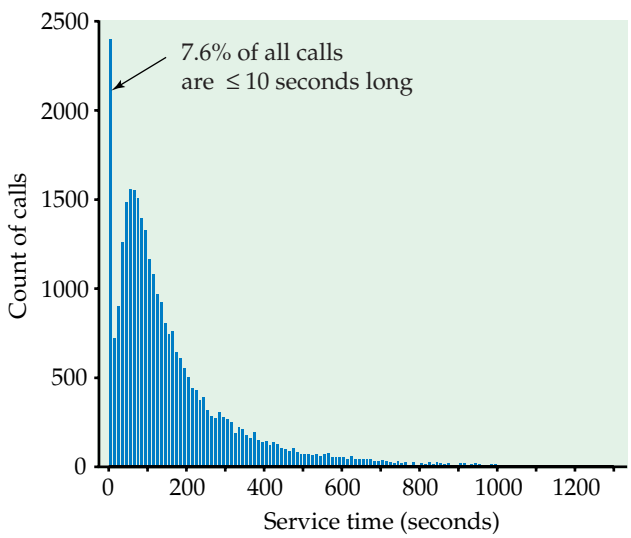


FIGURE 1.4 The distribution of call lengths for 31,492 calls to a bank's customer service center, for Example 1.6. The data show a surprising number of very short calls. These are mostly due to representatives deliberately hanging up in order to bring down their average call length.

tails

The extreme values of a distribution are in the **tails** of the distribution. The high values are in the upper, or right, tail and the low values are in the lower, or left, tail. The overall pattern in Figure 1.4 is made up of the many moderate call lengths and the long right tail of more lengthy calls. The striking departure from the overall pattern is the surprising number of very short calls in the left tail.

Our examination of the call center data illustrates some important principles:

- After you understand the background of your data (individuals, variables, units of measurement), the first thing to do is almost always **plot your data.**
- When you look at a plot, look for an **overall pattern** and also for any **striking departures** from the pattern.

We now turn to the kinds of graphs that are used to describe the distribution of a quantitative variable. We will explain how to make the graphs by hand, because knowing this helps you understand what the graphs show. However, making graphs by hand is so tedious that software is almost essential for effective data analysis unless you have just a few observations.

Stemplots

A *stemplot* (also called a stem-and-leaf plot) gives a quick picture of the shape of a distribution while including the actual numerical values in the graph. Stemplots work best for small numbers of observations that are all greater than 0.

STEMPLOT

To make a **stemplot:**

1. Separate each observation into a **stem** consisting of all but the final (rightmost) digit and a **leaf,** the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.

2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.

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

EXAMPLE

1.7 Literacy of men and women. The Islamic world is attracting increased attention in Europe and North America. Table 1.2 shows the percent of men and women at least 15 years old who were literate in 2002 in the major Islamic nations. We omitted countries with populations less than 3 million. Data for a few nations, such as Afghanistan and Iraq, are not available.[5]

To make a stemplot of the percents of females who are literate, use the first digits as stems and the second digits as leaves. Algeria's 60% literacy rate, for example, appears as the leaf 0 on the stem 6. Figure 1.5 shows the steps in making the plot.

TABLE 1.2

Literacy rates (percent) in Islamic nations

Country	Female percent	Male percent	Country	Female percent	Male percent
Algeria	60	78	Morocco	38	68
Bangladesh	31	50	Saudi Arabia	70	84
Egypt	46	68	Syria	63	89
Iran	71	85	Tajikistan	99	100
Jordan	86	96	Tunisia	63	83
Kazakhstan	99	100	Turkey	78	94
Lebanon	82	95	Uzbekistan	99	100
Libya	71	92	Yemen	29	70
Malaysia	85	92			

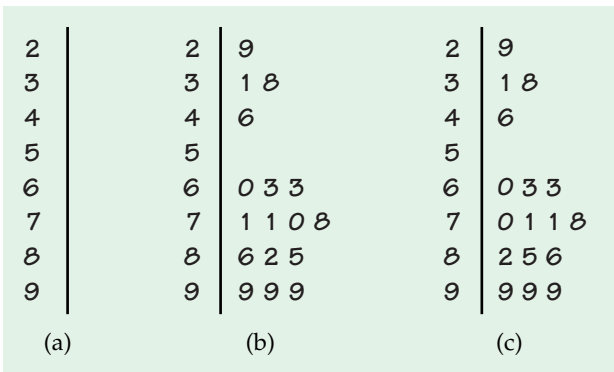


FIGURE 1.5 Making a stemplot of the data in Example 1.7. (a) Write the stems. (b) Go through the data and write each leaf on the proper stem. For example, the values on the 8 stem are 86, 82, and 85 in the order of the table. (c) Arrange the leaves on each stem in order out from the stem. The 8 stem now has leaves 2 5 6.

cluster

The overall pattern of the stemplot is irregular, as is often the case when there are only a few observations. There do appear to be two **clusters** of countries. The plot suggests that we might ask what explains the variation in literacy. For example, why do the three central Asian countries (Kazakhstan, Tajikistan, and Uzbekistan) have very high literacy rates?

USE YOUR KNOWLEDGE

1.5 **Make a stemplot.** Here are the scores on the first exam in an introductory statistics course for 30 students in one section of the course:

80	73	92	85	75	98	93	55	80	90	92	80	87	90	72
65	70	85	83	60	70	90	75	75	58	68	85	78	80	93

Use these data to make a stemplot. Then use the stemplot to describe the distribution of the first-exam scores for this course.

back-to-back stemplot

When you wish to compare two related distributions, a **back-to-back stemplot** with common stems is useful. The leaves on each side are ordered out from the common stem. Here is a back-to-back stemplot comparing the distributions of female and male literacy rates in the countries of Table 1.2.

Female		Male
9	2	
81	3	
6	4	
	5	0
330	6	88
8110	7	08
652	8	3459
999	9	22456
	10	000

The values on the left are the female percents, as in Figure 1.5, but ordered out from the stem from right to left. The values on the right are the male percents. It is clear that literacy is generally higher among males than among females in these countries.

Stemplots do not work well for large data sets, where each stem must hold a large number of leaves. Fortunately, there are two modifications of the basic stemplot that are helpful when plotting the distribution of a moderate number of observations. You can double the number of stems in a plot by **splitting each stem** into two: one with leaves 0 to 4 and the other with leaves 5 through 9. When the observed values have many digits, it is often best to **trim** the numbers by removing the last digit or digits before making a stemplot. You must use your judgment in deciding whether to split stems and whether to trim, though statistical software will often make these choices for you. Remember that the purpose of a stemplot is to display the shape of a distribution. If a stemplot has fewer than about five stems, you should usually split the stems unless there are few observations. If there are many stems with no leaves or only one leaf, trimming will reduce the number of stems. Here is an example that makes use of both of these modifications.

splitting stems
trimming

EXAMPLE

1.8 Stemplot for length of service calls. Return to the 80 customer service call lengths in Table 1.1. To make a stemplot of this distribution, we first trim the call lengths to tens of seconds by dropping the last digit. For example, 56 seconds trims to 5 and 143 seconds trims to 14. (We might also round to the nearest 10 seconds, but trimming is faster than rounding if you must do it by hand.)

We can then use tens of seconds as our leaves, with the digits to the left forming stems. This gives us the single-digit leaves that a stemplot requires. For example, 56 trimmed to 5 becomes leaf 5 on the 0 stem; 143 trimmed to 14 becomes leaf 4 on the 1 stem.

Because we have 80 observations, we split the stems. Thus, 56 trimmed to 5 becomes leaf 5 on the second 0 stem, along with all leaves 5 to 9. Leaves

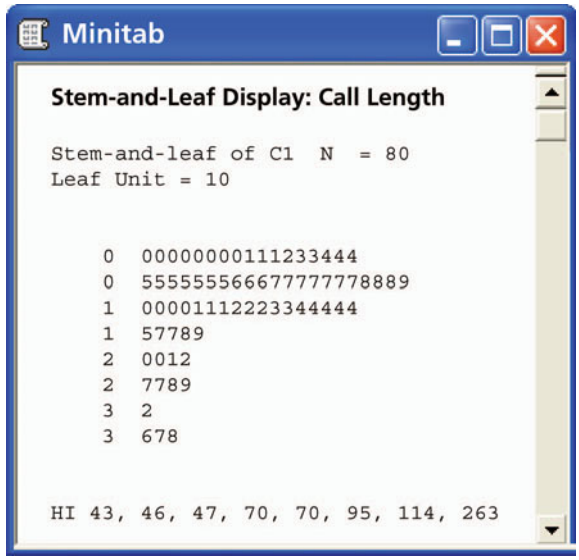


FIGURE 1.6 Stemplot from Minitab of the 80 call lengths in Table 1.1, for Example 1.8. The software has trimmed the data by removing the last digit. It has also split stems and listed the highest observations apart from the plot.

0 to 4 go on the first 0 stem. Figure 1.6 is a stemplot of these data made by software. The software automatically did what we suggest: trimmed to tens of seconds and split stems. To save space, the software also listed the largest values as "HI" rather than create stems all the way up to 26. The stemplot shows the overall pattern of the distribution, with many short to moderate lengths and some very long calls.

Histograms

Stemplots display the actual values of the observations. This feature makes stemplots awkward for large data sets. Moreover, the picture presented by a stemplot divides the observations into groups (stems) determined by the number system rather than by judgment. Histograms do not have these limitations.

histogram

A **histogram** breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should always choose classes of equal width. Histograms are slower to construct by hand than stemplots and do not display the actual values observed. For these reasons we prefer stemplots for small data sets. The construction of a histogram is best shown by example. Any statistical software package will of course make a histogram for you.

EXAMPLE

1.9 Distribution of IQ scores. You have probably heard that the distribution of scores on IQ tests is supposed to be roughly "bell-shaped." Let's look at some actual IQ scores. Table 1.3 displays the IQ scores of 60 fifth-grade students chosen at random from one school.[6]

1. Divide the range of the data into classes of equal width. The scores in Table 1.3 range from 81 to 145, so we choose as our classes