HHS HIGHLAND HIGH SCHOOL

Antelope Valley Union High School District

Bulldogs

TO: Prospective AP Statistics Students
FROM: Mrs. Arispe
RE: SUMMER ASSIGNMENT FOR AP STATISTICS

Thank you for signing up to take an AP course and in particular AP STATS! I am looking forward to again teaching AP Statistics this upcoming school year. This class often feels like it is easy regarding the mathematical procedures as most are completely do-able via a TI-84 graphing calculator. However, there is quite a bit of writing and explanation of the purpose behind the use of the procedure, therein lies the greatest difficulty for most.

The following pages of the fundamentals you will need for class; most of which was covered in previous math courses. Please do this packet in a spiral notebook, which will be collected on the second day of class. It is recommended that you use the graph paper spiral because you will be doing many graphs and pictures. You need to take notes on ALL sections and COMPLETE #1-35 ALL. You do NOT need to write out each problem but please write out the title of each.

You are **REQUIRED** to have a graphics display calculator. The book we will be using is based on the **TI-84 Plus CE**, so that is what I HIGHLY recommend; I am familiar with this calculator as well. You can find good deals during the summer months prior to the beginning of school when there are sales. You may also want to look online as many students may sell theirs right after testing.

The above assignment will be due the 2nd day of class and there will be a quiz within the first 10 days on this material. So try to enjoy your summer, get your work done, and get ready to read and explain what the meaning of the data is. It should be a fun but challenging year.

Peace out,
Paulette Arispe
AP STATISTICS
parispe@avhsd.org

chapter 1

Exploring Data

ACTIVITY 1 How Fast Is Your Heart Beating?

Materials: Clock or watch with second hand

A person's pulse rate provides information about the health of his or her heart. Would you expect to find a difference between male and female pulse rates? In this activity, you and your classmates will collect some data to try to answer this question.

1. To determine your pulse rate, hold the *fingers* of one hand on the artery in your neck or on the inside of the wrist. (The thumb should not be used, because there is a pulse in the thumb.) Count the number of pulse beats in one minute. Do this three times, and calculate your *average* individual pulse rate (add your three pulse rates and divide by 3.) Why is doing this three times better than doing it once?
2. Record the pulse rates for the class in a table, with one column for males and a second column for females. Are there any unusual pulse rates?
3. For now, simply calculate the average pulse rate for the males and the average pulse rate for the females, and compare.

INTRODUCTION

Statistics is the science of data. We begin our study of statistics by mastering the art of examining data. 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 by a set of data. Individuals may be people, but they may also be animals or things.

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

A college's student data base, for example, includes data about every currently enrolled student. The students are the *individuals* described by the data set. For each individual, the data contain the values of *variables* such as age, gender (female or male), choice of major, and grade point average. In practice, any set of data is accompanied by background information that helps us understand the data.

When you meet a new set of data, ask yourself the following questions:

1. **Who?** What **individuals** do the data describe? **How many** individuals appear in the data?

2. **What?** How many **variables** are there? What are the **exact definitions** of these variables? In what **units** is each variable recorded? Weights, for example, might be recorded in pounds, in thousands of pounds, or in kilograms. Is there any reason to mistrust the values of any variable?

3. **Why?** What is the reason the data were gathered? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones we actually have data for?

Some variables, like gender and college major, simply place individuals into categories. Others, like age and grade point average (GPA), take numerical values for which we can do arithmetic. It makes sense to give an average GPA for a college's students, but it does not make sense to give an "average" gender. We can, however, count the numbers of female and male students and do arithmetic with these counts.

CATEGORICAL AND QUANTITATIVE VARIABLES

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

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

EXAMPLE 1.1 EDUCATION IN THE UNITED STATES

Here is a small part of a data set that describes public education in the United States:

State	Region	Population (1000)	SAT Verbal	SAT Math	Percent taking	Percent no HS	Teachers' pay ($1000)
⋮							
CA	PAC	33,871	497	514	49	23.8	43.7
CO	MTN	4,301	536	540	32	15.6	37.1
CT	NE	3,406	510	509	80	20.8	50.7
⋮							

Let's answer the three "W" questions about these data.

1. **Who?** The *individuals* described are the states. There are 51 of them, the 50 states and the District of Columbia, but we give data for only 3. Each row in the table describes one individual. You will often see each row of data called a ***case.***

case

2. **What?** Each column contains the values of one variable for all the individuals. This is the usual arrangement in data tables. Seven variables are recorded for each state. The first column identifies the state by its two-letter post office code. We give data for California, Colorado, and Connecticut. The second column says which region of the country the state is in. The Census Bureau divides the nation into nine regions. These three are Pacific, Mountain, and New England. The third column contains state populations, in thousands of people. Be sure to notice that the *units* are thousands of people. California's 33,871 stands for 33,871,000 people. The population data come from the 2000 census. They are therefore quite accurate as of April 1, 2000, but don't show later changes in population.

The remaining five variables are the average scores of the states' high school seniors on the SAT verbal and mathematics exams, the percent of seniors who take the SAT, the percent of students who did not complete high school, and average teachers' salaries in thousands of dollars. Each of these variables needs more explanation before we can fully understand the data.

3. **Why?** Some people will use these data to evaluate the quality of individual states' educational programs. Others may compare states on one or more of the variables. Future teachers might want to know how much they can expect to earn.

A variable generally takes values that vary. One variable may take values that are very close together while another variable takes values that are quite spread out. We say that the *pattern of variation* of a variable is its *distribution.*

DISTRIBUTION

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

exploratory data analysis

Statistical tools and ideas can help you examine data in order to describe their main features. This examination is called ***exploratory data analysis.*** Like an explorer crossing unknown lands, we first simply describe what we see. Each example we meet will have some background information to help us, but our emphasis is on examining the data. 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 relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will organize our learning the same way. Chapters 1 and 2 examine single-variable data, and Chapters 3 and 4 look at relationships among variables. In both settings, we begin with graphs and then move on to numerical summaries.

EXERCISES

1.1 **FUEL-EFFICIENT CARS** Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 1998 model motor vehicles:

Make and Model	Vehicle type	Transmission type	Number of cylinders	City MPG	Highway MPG
⋮					
BMW 318I	Subcompact	Automatic	4	22	31
BMW 318I	Subcompact	Manual	4	23	32
Buick Century	Midsize	Automatic	6	20	29
Chevrolet Blazer	Four-wheel drive	Automatic	6	16	20
⋮					

(a) What are the individuals in this data set?

(b) For each individual, what variables are given? Which of these variables are categorical and which are quantitative?

1.2 **MEDICAL STUDY VARIABLES** Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?

(a) Gender (female or male)

(b) Age (years)

(c) Race (Asian, black, white, or other)

(d) Smoker (yes or no)

(e) Systolic blood pressure (millimeters of mercury)

(f) Level of calcium in the blood (micrograms per milliliter)

1.3 You want to compare the "size" of several statistics textbooks. Describe at least three possible numerical variables that describe the "size" of a book. In what *units* would you measure each variable?

1.4 Popular magazines often rank cities in terms of how desirable it is to live and work in each city. Describe five variables that you would measure for each city if you were designing such a study. Give reasons for each of your choices.

1.1 DISPLAYING DISTRIBUTIONS WITH GRAPHS

Displaying categorical variables: bar graphs and pie charts

The values of a categorical variable are labels for the categories, such as "male" and "female." The distribution of a categorical variable lists the categories and gives either the **count** or the **percent** of individuals who fall in each category.

EXAMPLE 1.2 THE MOST POPULAR SOFT DRINK

The following table displays the sales figures and market share (percent of total sales) achieved by several major soft drink companies in 1999. That year, a total of 9930 million cases of soft drink were sold.[1]

Company	Cases sold (millions)	Market share (percent)
Coca-Cola Co.	4377.5	44.1
Pepsi-Cola Co.	3119.5	31.4
Dr. Pepper/7-Up (Cadbury)	1455.1	14.7
Cott Corp.	310.0	3.1
National Beverage	205.0	2.1
Royal Crown	115.4	1.2
Other	347.5	3.4

How to construct a bar graph:

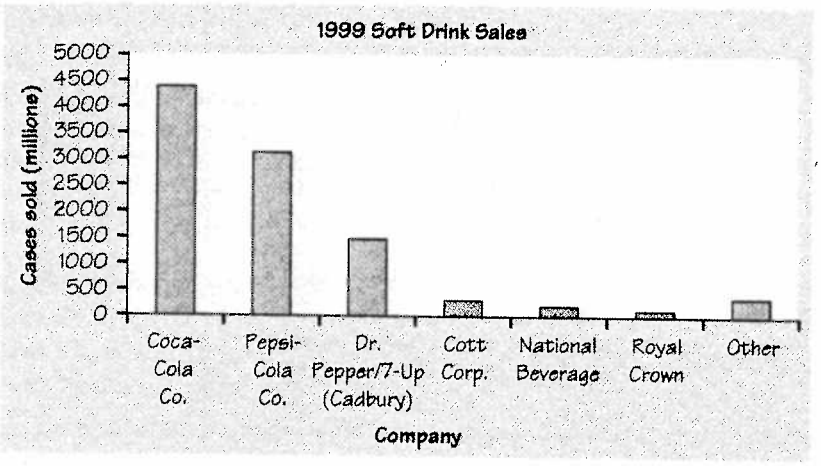
Step 1: Label your axes and title your graph. Draw a set of axes. Label the horizontal axis "Company" and the vertical axis "Cases sold." Title your graph.

Step 2: Scale your axes. Use the counts in each category to help you scale your vertical axis. Write the category names at equally spaced intervals beneath the horizontal axis.

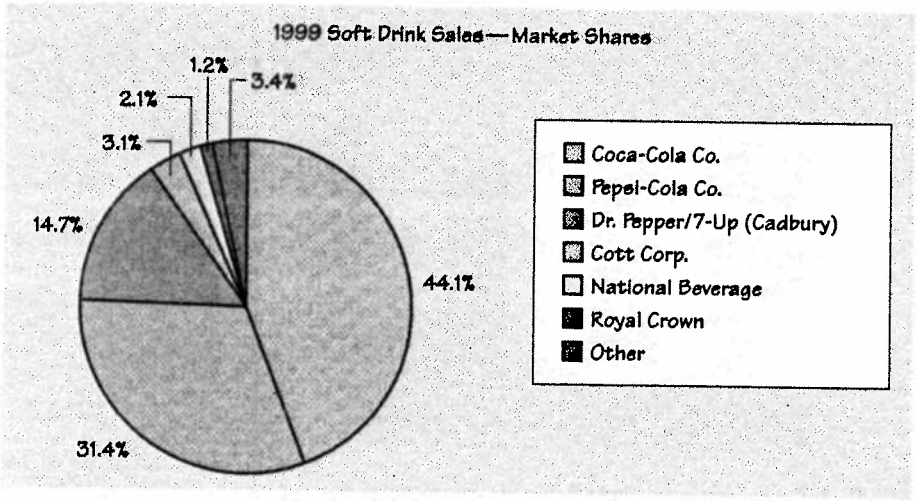
Step 3: Draw a vertical bar above each category name to a height that corresponds to the count in that category. For example, the height of the "Pepsi-Cola Co." bar should be at 3119.5 on the vertical scale. *Leave a space between the bars in a bar graph.*

Figure 1.1(a) displays the completed bar graph.

How to construct a pie chart: Use a computer! Any statistical software package and many spreadsheet programs will construct these plots for you. Figure 1.1(b) is a pie chart for the soft drink sales data.



(a)



(b)

FIGURE 1.1 A bar graph (a) and a pie chart (b) displaying soft drink sales by companies in 1999.

The **bar graph** in Figure 1.1(a) quickly compares the soft drink sales of the companies. The heights of the bars show the counts in the seven categories. The **pie chart** in Figure 1.1(b) helps us see what part of the whole each group forms. For example, the Coca-Cola "slice" makes up 44.1% of the pie because the Coca-Cola Company sold 44.1% of all soft drinks in 1999.

Bar graphs and pie charts help an audience grasp the distribution quickly. To make a pie chart, you must include all the categories that make up a whole. Bar graphs are more flexible.

EXAMPLE 1.3 DO YOU WEAR YOUR SEAT BELT?

In 1998, the National Highway and Traffic Safety Administration (NHTSA) conducted a study on seat belt use. The table below shows the percentage of automobile drivers who were observed to be wearing their seat belts in each region of the United States.[2]

Region	Percent wearing seat belts
Northeast	66.4
Midwest	63.6
South	78.9
West	80.8

Figure 1.2 shows a bar graph for these data. Notice that the vertical scale is measured in percents.

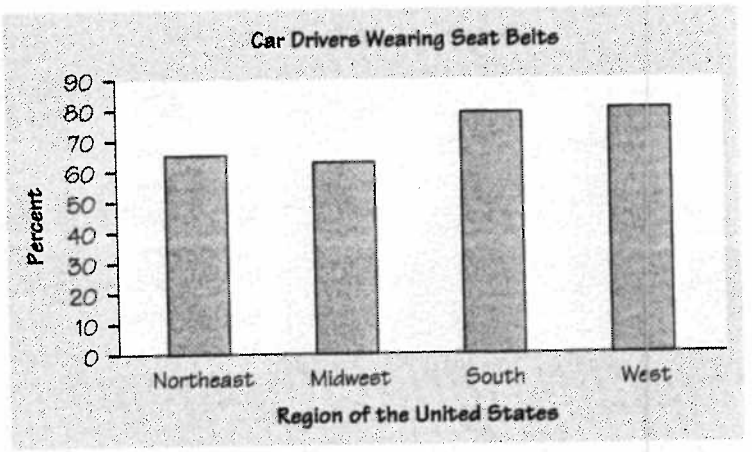


FIGURE 1.2 A bar graph showing the percentage of drivers who wear their seat belts in each of four U.S. regions.

Drivers in the South and West seem to be more concerned about wearing seat belts than those in the Northeast and Midwest. It is not possible to display these data in a single pie chart, because the four percentages cannot be combined to yield a whole (their sum is well over 100%).

EXERCISES

1.5 **FEMALE DOCTORATES** Here are data on the percent of females among people earning doctorates in 1994 in several fields of study:[3]

Computer science	15.4%	Life sciences	40.7%
Education	60.8%	Physical sciences	21.7%
Engineering	11.1%	Psychology	62.2%

(a) Present these data in a well-labeled bar graph.

(b) Would it also be correct to use a pie chart to display these data? If so, construct the pie chart. If not, explain why not.

1.6 **ACCIDENTAL DEATHS** In 1997 there were 92,353 deaths from accidents in the United States. Among these were 42,340 deaths from motor vehicle accidents, 11,858 from falls, 10,163 from poisoning, 4051 from drowning, and 3601 from fires.[4]

(a) Find the percent of accidental deaths from each of these causes, rounded to the nearest percent. What percent of accidental deaths were due to other causes?

(b) Make a well-labeled bar graph of the distribution of causes of accidental deaths. Be sure to include an "other causes" bar.

(c) Would it also be correct to use a pie chart to display these data? If so, construct the pie chart. If not, explain why not.

Displaying quantitative variables: dotplots and stemplots

Several types of graphs can be used to display quantitative data. One of the simplest to construct is a **dotplot**.

EXAMPLE 1.4 GOOOOOOOOAAAAALLLLLLLL!!!

The number of goals scored by each team in the first round of the California Southern Section Division V high school soccer playoffs is shown in the following table.[5]

5	0	1	0	7	2	1	0	4	0	3	0	2	0
3	1	5	0	3	0	1	0	1	0	2	0	3	1

How to construct a dotplot:

Step 1: Label your axis and title your graph. Draw a horizontal line and label it with the variable (in this case, number of goals scored). Title your graph.

Step 2: Scale the axis based on the values of the variable.

Step 3: Mark a dot above the number on the horizontal axis corresponding to each data value. Figure 1.3 displays the completed dotplot.

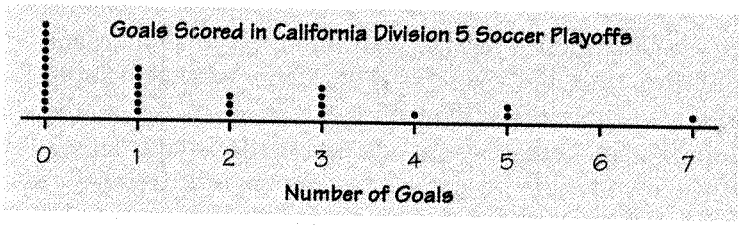


FIGURE 1.3 Goals scored by teams in the California Southern Section Division V high school soccer playoffs.

Making a statistical graph is not an end in itself. After all, a computer or graphing calculator can make graphs faster than we can. The purpose of the graph is to help us understand the data. After you (or your calculator) make a graph, always ask, "What do I see?" Here is a general tactic for looking at graphs: *Look for an overall pattern and also for striking deviations from that pattern.*

OVERALL PATTERN OF A DISTRIBUTION

To describe the overall pattern of a distribution:

- Give the **center** and the **spread**.
- See if the distribution has a simple **shape** that you can describe in a few words.

Section 1.2 tells in detail how to measure center and spread. For now, describe the *center* by finding a value that divides the observations so that about half take larger values and about half have smaller values. In Figure 1.3, the center is 1. That is, a typical team scored about 1 goal in its playoff soccer game. You can describe the *spread* by giving the smallest and largest values. The spread in Figure 1.3 is from 0 goals to 7 goals scored.

The dotplot in Figure 1.3 shows that in most of the playoff games, Division V soccer teams scored very few goals. There were only four teams that scored 4 or more goals. We can say that the distribution has a "long tail" to the right, or that its *shape* is "skewed right." You will learn more about describing shape shortly.

Is the one team that scored 7 goals an *outlier*? This value certainly differs from the overall pattern. To some extent, deciding whether an observation is an outlier is a matter of judgment. We will introduce an objective criterion for determining outliers in Section 1.2.

OUTLIERS

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

Once you have spotted outliers, look for an explanation. Many outliers are due to mistakes, such as typing 4.0 as 40. Other outliers point to the special nature of some observations. Explaining outliers usually requires some background information. Perhaps the soccer team that scored seven goals has some very talented offensive players. Or maybe their opponents played poor defense.

Sometimes the values of a variable are too spread out for us to make a reasonable dotplot. In these cases, we can consider another simple graphical display: a **stemplot.**

EXAMPLE 1.5 WATCH THAT CAFFEINE!

The U.S. Food and Drug Administration limits the amount of caffeine in a 12-ounce can of carbonated beverage to 72 milligrams (mg). Data on the caffeine content of popular soft drinks are provided in Table 1.1. How does the caffeine content of these drinks compare to the USFDA's limit?

TABLE 1.1 Caffeine content (in milligrams) for an 8-ounce serving of popular soft drinks

Brand	Caffeine (mg per 8-oz. serving)	Brand	Caffeine (mg per 8-oz. serving)
A&W Cream Soda	20	IBC Cherry Cola	16
Barq's root beer	15	Kick	38
Cherry Coca-Cola	23	KMX	36
Cherry RC Cola	29	Mello Yello	35
Coca-Cola Classic	23	Mountain Dew	37
Diet A&W Cream Soda	15	Mr. Pibb	27
Diet Cherry Coca-Cola	23	Nehi Wild Red Soda	33
Diet Coke	31	Pepsi One	37
Diet Dr. Pepper	28	Pepsi-Cola	25
Diet Mello Yello	35	RC Edge	47
Diet Mountain Dew	37	Red Flash	27
Diet Mr. Pibb	27	Royal Crown Cola	29
Diet Pepsi-Cola	24	Ruby Red Squirt	26
Diet Ruby Red Squirt	26	Sun Drop Cherry	43
Diet Sun Drop	47	Sun Drop Regular	43
Diet Sunkist Orange Soda	28	Sunkist Orange Soda	28
Diet Wild Cherry Pepsi	24	Surge	35
Dr. Nehi	28	TAB	31
Dr. Pepper	28	Wild Cherry Pepsi	25

Source: National Soft Drink Association, 1999.

The caffeine levels spread from 15 to 47 milligrams for these soft drinks. You could make a dotplot for these data, but a stemplot might be preferable due to the large spread.

How to construct a stemplot:

Step 1: Separate each observation into a *stem* consisting of all but the rightmost digit and a *leaf*, the final digit. A&W Cream Soda has 20 milligrams of caffeine per 8-ounce serving. The number 2 is the stem and 0 is the leaf.

Step 2: Write the stems vertically in increasing order from top to bottom, and draw a vertical line to the right of the stems. Go through the data, writing each leaf to the right of its stem and spacing the leaves equally.

```
1 | 5 5 6
2 | 0 3 9 3 3 8 7 4 6 8 4 8 8 7 5 7 9 6 8 5
3 | 1 5 7 8 6 5 7 3 7 5 1
4 | 7 7 3 3
```

Step 3: Write the stems again, and rearrange the leaves in increasing order out from the stem.

Step 4: Title your graph and add a key describing what the stems and leaves represent. Figure 1.4(a) shows the completed stemplot.

What *shape* does this distribution have? It is difficult to tell with so few stems. We can get a better picture of the caffeine content in soft drinks by "splitting stems." In Figure 1.4(a), the values from 10 to 19 milligrams are placed on the "1" stem. Figure 1.4(b) shows another stemplot of the same data. This time, values having leaves 0 through 4 are placed on one stem, while values ending in 5 through 9 are placed on another stem.

Now the bimodal (two-peaked) *shape* of the distribution is clear. Most soft drinks seem to have between 25 and 29 milligrams or between 35 and 38 milligrams of caffeine per 8-ounce serving. The center of the distribution is 28 milligrams per 8-ounce serving. At first glance, it looks like none of these soft drinks even comes close to the USFDA's caffeine limit of 72 milligrams per 12-ounce serving. Be careful! The values in the stemplot are given in milligrams per 8-ounce serving. Two soft drinks have caffeine levels of 47 milligrams per 8-ounce serving. A 12-ounce serving of these beverages would have 1.5(47) = 70.5 milligrams of caffeine. Always check the units of measurement!

CAFFEINE CONTENT (MG) PER 8-OUNCE SERVING OF VARIOUS SOFT DRINKS

```
1 | 5 5 6
2 | 0 3 3 3 4 4 5 5 6 6 7 7 7 8 8 8 8 8 9 9
3 | 1 1 3 5 5 5 5 6 7 7 7 8
4 | 3 3 7 7
```

(a)

Key:
3|5 means the soft drink contains 35 mg of caffeine per 8-ounce serving.

```
1 | 5 5 6
2 | 0 3 3 3 4 4
2 | 5 5 6 6 7 7 7 8 8 8 8 8 9 9
3 | 1 1 3
3 | 5 5 5 6 7 7 7 8
4 | 3 3
4 | 7 7
```

(b)

Key:
2|8 means the soft drink contains 28 mg of caffeine per 8-ounce serving.

FIGURE 1.4 Two stemplots showing the caffeine content (mg) of various soft drinks. Figure 1.4(b) improves on the stemplot of Figure 1.4(a) by splitting stems.

Here are a few tips for you to consider when you want to construct a stemplot:

- Whenever you split stems, be sure that each stem is assigned an equal number of possible leaf digits.
- There is no magic number of stems to use. Too few stems will result in a skyscraper-shaped plot, while too many stems will yield a very flat "pancake" graph.

- Five stems is a good minimum.
- You can get more flexibility by *rounding* the data so that the final digit after rounding is suitable as a leaf. Do this when the data have too many digits.

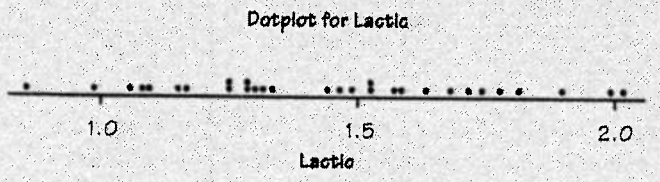
The chief advantages of dotplots and stemplots are that they are easy to construct and they display the actual data values (unless we round). Neither will work well with large data sets. Most statistical software packages will make dotplots and stemplots for you. That will allow you to spend more time making sense of the data.

TECHNOLOGY TOOLBOX *Interpreting computer output*

As cheddar cheese matures, a variety of chemical processes take place. The taste of mature cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the Latrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition. The final concentrations of lactic acid in the 30 samples, as a multiple of their initial concentrations, are given below.[6]

A dotplot and a stemplot from the Minitab statistical software package are shown in Figure 1.5. The dots in the dotplot are so spread out that the distribution seems to have no distinct shape. The stemplot does a better job of summarizing the data.

0.86	1.53	1.57	1.81	0.99	1.09	1.29	1.78	1.29	1.58
1.68	1.90	1.06	1.30	1.52	1.74	1.16	1.49	1.63	1.99
1.15	1.33	1.44	2.01	1.31	1.46	1.72	1.25	1.08	1.25




```
Stem-and-leaf of Lactic      N = 30
Leaf Unit = 0.010

 1   8   6
 2   9   9
 5  10   689
 7  11   56
11  12   5599
14  13   013
(3) 14   469
13  15   2378
 9  16   38
 7  17   248
 4  18   1
 3  19   09
 1  20   1
```

FIGURE 1.5 Minitab dotplot and stemplot for cheese data.

TECHNOLOGY TOOLBOX *Interpreting computer output (continued)*

Notice how the data are recorded in the stemplot. The "leaf unit" is 0.01, which tells us that the stems are given in tenths and the leaves are given in hundredths. We can see that the *spread* of the lactic acid concentrations is from 0.86 to 2.01. Where is the *center* of the distribution? Minitab counts the number of observations from the bottom up and from the top down and lists those counts to the left of the stemplot. Since there are 30 observations, the "middle value" would fall between the 15th and 16th data values from either end—at 1.45. The (3) to the far left of this stem is Minitab's way of marking the location of the "middle value." So a typical sample of mature cheese has 1.45 times as much lactic acid as it did initially. The distribution is roughly symmetrical in *shape*. There appear to be no *outliers*.

EXERCISES

1.7 OLYMPIC GOLD Athletes like Cathy Freeman, Rulon Gardner, Ian Thorpe, Marion Jones, and Jenny Thompson captured public attention by winning gold medals in the 2000 Summer Olympic Games in Sydney, Australia. Table 1.2 displays the total number of gold medals won by several countries in the 2000 Summer Olympics.

TABLE 1.2 Gold medals won by selected countries in the 2000 Summer Olympics

Country	Gold medals	Country	Gold medals
Sri Lanka	0	Netherlands	12
Qatar	0	India	0
Vietnam	0	Georgia	0
Great Britain	28	Kyrgyzstan	0
Norway	10	Costa Rica	0
Romania	26	Brazil	0
Switzerland	9	Uzbekistan	1
Armenia	0	Thailand	1
Kuwait	0	Denmark	2
Bahamas	1	Latvia	1
Kenya	2	Czech Republic	2
Trinidad and Tobago	0	Hungary	8
Greece	13	Sweden	4
Mozambique	1	Uruguay	0
Kazakhstan	3	United States	39

Source: BBC Olympics Web site.

Make a dotplot to display these data. Describe the distribution of number of gold medals won.

1.8 ARE YOU DRIVING A GAS GUZZLER? Table 1.3 displays the highway gas mileage for 32 model year 2000 midsize cars.

TABLE 1.3 Highway gas mileage for model year 2000 midsize cars

Model	MPG	Model	MPG
Acura 3.5RL	24	Lexus GS300	24
Audi A6 Quattro	24	Lexus LS400	25
BMW 740I Sport M	21	Lincoln-Mercury LS	25
Buick Regal	29	Lincoln-Mercury Sable	28
Cadillac Catera	24	Mazda 626	28
Cadillac Eldorado	28	Mercedes-Benz E320	30
Chevrolet Lumina	30	Mercedes-Benz E430	24
Chrysler Cirrus	28	Mitsubishi Diamante	25
Dodge Stratus	28	Mitsubishi Galant	28
Honda Accord	29	Nissan Maxima	28
Hyundai Sonata	28	Oldsmobile Intrigue	28
Infiniti I30	28	Saab 9-3	26
Infiniti Q45	23	Saturn LS	32
Jaguar Vanden Plas	24	Toyota Camry	30
Jaguar S/C	21	Volkswagon Passat	29
Jaguar X200	26	Volvo S70	27

(a) Make a dotplot of these data.

(b) Describe the shape, center, and spread of the distribution of gas mileages. Are there any potential outliers?

1.9 **MICHIGAN COLLEGE TUITIONS** There are 81 colleges and universities in Michigan. Their tuition and fees for the 1999 to 2000 school year run from $1260 at Kalamazoo Valley Community College to $19,258 at Kalamazoo College. Figure 1.6 (next page) shows a stemplot of the tuition charges.

(a) What do the stems and leaves represent in the stemplot? Have the data been rounded?

(b) Describe the shape, center, and spread of the tuition distribution. Are there any outliers?

1.10 **DRP TEST SCORES** There are many ways to measure the reading ability of children. One frequently used test is the Degree of Reading Power (DRP). In a research study on third-grade students, the DRP was administered to 44 students.[7] Their scores were:

40	26	39	14	42	18	25	43	46	27	19
47	19	26	35	34	15	44	40	38	31	46
52	25	35	35	33	29	34	41	49	28	52
47	35	48	22	33	41	51	27	14	54	45

Display these data graphically. Write a paragraph describing the distribution of DRP scores.

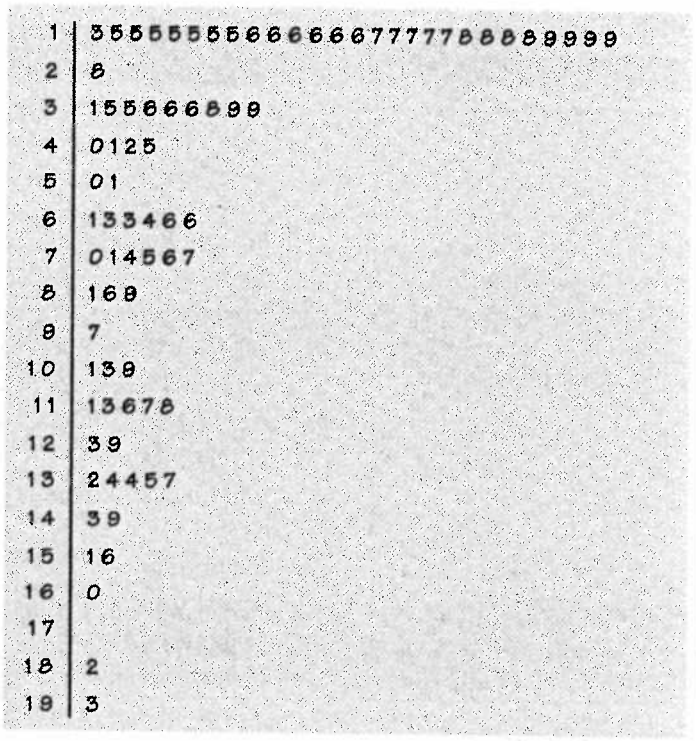


FIGURE 1.6 Stemplot of the Michigan tuition and fee data, for Exercise 1.9.

1.11 SHOPPING SPREE! A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (in dollars), arranged in increasing order:

3.11	8.88	9.26	10.81	12.69	13.78	15.23	15.62	17.00	17.39
18.36	18.43	19.27	19.50	19.54	20.16	20.59	22.22	23.04	24.47
24.58	25.13	26.24	26.26	27.65	28.06	28.08	28.38	32.03	34.98
36.37	38.64	39.16	41.02	42.97	44.08	44.67	45.40	46.69	48.65
50.39	52.75	54.80	59.07	61.22	70.32	82.70	85.76	86.37	93.34

(a) Round each amount to the nearest dollar. Then make a stemplot using tens of dollars as the stem and dollars as the leaves.

(b) Make another stemplot of the data by splitting stems. Which of the plots shows the shape of the distribution better?

(c) Describe the shape, center, and spread of the distribution. Write a few sentences describing the amount of money spent by shoppers at this supermarket.

Displaying quantitative variables: histograms

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

EXAMPLE 1.6 PRESIDENTIAL AGES AT INAUGURATION

How old are presidents at their inaugurations? Was Bill Clinton, at age 46, unusually young? Table 1.4 gives the data, the ages of all U.S presidents when they took office.

TABLE 1.4 Ages of the presidents at inauguration

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. D. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	61
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. B. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G. Bush	64
Taylor	64	Taft	51	Clinton	46
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55		
Buchanan	65	Coolidge	51		

How to make a histogram:

Step 1: Divide the range of the data into classes of equal width. Count the number of observations in each class. The data in Table 1.4 range from 42 to 69, so we choose as our classes

$$40 \le \text{president's age at inauguration} < 45$$
$$45 \le \text{president's age at inauguration} < 50$$
$$\vdots$$
$$65 \le \text{president's age at inauguration} < 70$$

Be sure to specify the classes precisely so that each observation falls into exactly one class. Martin Van Buren, who was age 54 at the time of his inauguration, would fall into the third class interval. Grover Cleveland, who was age 55, would be placed in the fourth class interval.

Here are the counts:

Class	Count
40–44	2
45–49	6
50–54	13
55–59	12
60–64	7
65–69	3

Step 2: Label and scale your axes and title your graph. Label the horizontal axis "Age at inauguration" and the vertical axis "Number of presidents." For the classes we chose, we should scale the horizontal axis from 40 to 70, with tick marks 5 units apart. The vertical axis contains the scale of counts and should range from 0 to at least 13.

Step 3: Draw a bar that represents the count in each class. The base of a bar should cover its class, and the bar height is the class count. Leave no horizontal space between the bars (unless a class is empty, so that its bar has height 0). Figure 1.7 shows the completed histogram.

Graphing note: It is common to add a "break-in-scale" symbol (//) on an axis that does not start at 0, like the horizontal axis in this example.

Interpretation:

Center: It appears that the typical age of a new president is about 55 years, because 55 is near the center of the histogram.

Spread: As the histogram in Figure 1.7 shows, there is a good deal of variation in the ages at which presidents take office. Teddy Roosevelt was the youngest, at age 42, and Ronald Reagan, at age 69, was the oldest.

Shape: The distribution is roughly symmetric and has a single peak (unimodal).

Outliers: There appear to be no outliers.

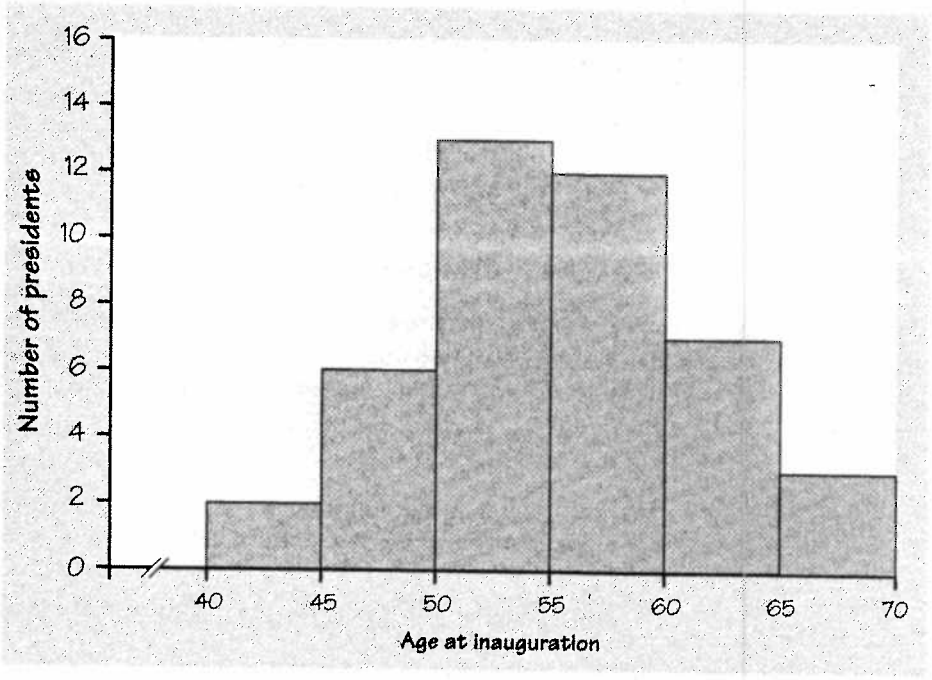


FIGURE 1.7 The distribution of the ages of presidents at their inaugurations, from Table 1.4.

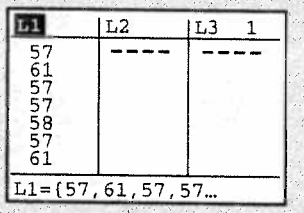
You can also use computer software or a calculator to construct histograms.

TECHNOLOGY TOOLBOX *Making calculator histograms*

1. Enter the presidential age data from Example 1.6 in your statistics list editor.

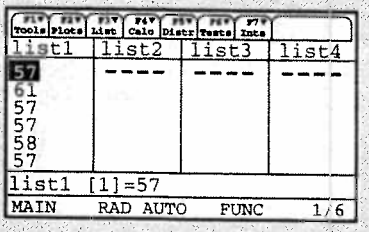
TI-83

- Press STAT and choose 1:Edit....
- Type the values into list L_1.



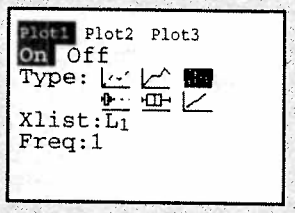
TI-89

- Press APPS, choose 1:FlashApps, then select Stats/List Editor and press ENTER.
- Type the values into list1.

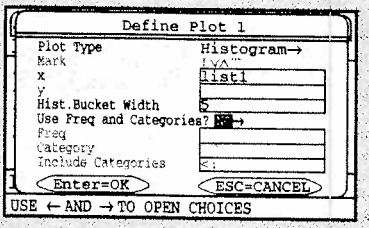


2. Set up a histogram in the statistics plots menu.

- Press 2nd Y= (STAT PLOT).
- Press ENTER to go into Plot1.
- Adjust your settings as shown.

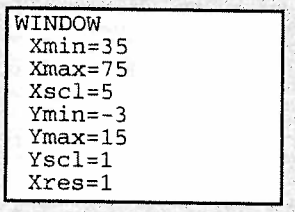


- Press F2 and choose 1:Plot Setup....
- With Plot 1 highlighted, press F1 to define.
- Change Hist. Bucket Width to 5, as shown.

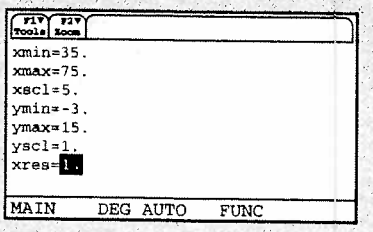


3. Set the window to match the class intervals chosen in Example 1.6.

- Press WINDOW.
- Enter the values shown.



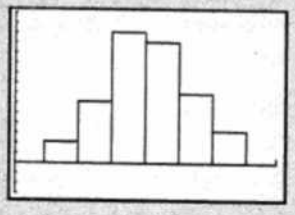
- Press ◆ F2 (WINDOW).
- Enter the values shown.

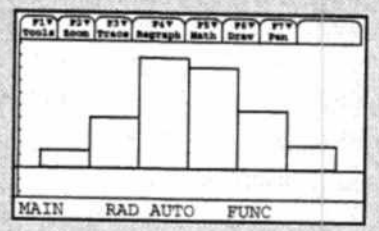


4. Graph the histogram. Compare with Figure 1.7.

- Press GRAPH.
- Press ◆ F3 (GRAPH).

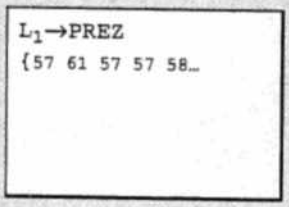
TECHNOLOGY TOOLBOX *Making calculator histograms (continued)*

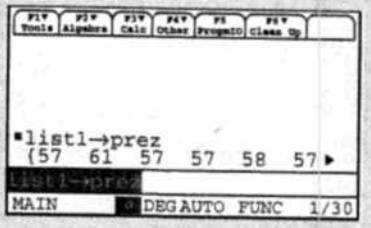




5. Save the data in a named list for later use.

- From the home screen, type the command `L1→PREZ` (`list1→prez` on the TI-89) and press ENTER. The data are now stored in a list called PREZ.





Histogram tips:

- There is no one right choice of the classes in a histogram. Too few classes will give a "skyscraper" graph, with all values in a few classes with tall bars. Too many will produce a "pancake" graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution.
- Five classes is a good minimum.
- Our eyes respond to the *area* of the bars in a histogram, so be sure to choose classes that are all the same width. Then area is determined by height and all classes are fairly represented.
- If you use a computer or graphing calculator, beware of letting the device choose the classes.

EXERCISES

1.12 WHERE DO OLDER FOLKS LIVE? Table 1.5 gives the percentage of residents aged 65 or older in each of the 50 states.

TABLE 1.5 Percent of the population in each state aged 65 or older

State	Percent	State	Percent	State	Percent
Alabama	13.1	Louisiana	11.5	Ohio	13.4
Alaska	5.5	Maine	14.1	Oklahoma	13.4
Arizona	13.2	Maryland	11.5	Oregon	13.2
Arkansas	14.3	Massachusetts	14.0	Pennsylvania	15.9
California	11.1	Michigan	12.5	Rhode Island	15.6
Colorado	10.1	Minnesota	12.3	South Carolina	12.2
Connecticut	14.3	Mississippi	12.2	South Dakota	14.3
Delaware	13.0	Missouri	13.7	Tennessee	12.5
Florida	18.3	Montana	13.3	Texas	10.1
Georgia	9.9	Nebraska	13.8	Utah	8.8
Hawaii	13.3	Nevada	11.5	Vermont	12.3
Idaho	11.3	New Hampshire	12.0	Virginia	11.3
Illinois	12.4	New Jersey	13.6	Washington	11.5
Indiana	12.5	New Mexico	11.4	West Virginia	15.2
Iowa	15.1	New York	13.3	Wisconsin	13.2
Kansas	13.5	North Carolina	12.5	Wyoming	11.5
Kentucky	12.5	North Dakota	14.4		

Source: U.S. Census Bureau, 1998.

(a) Construct a histogram to display these data. Record your class intervals and counts.

(b) Describe the distribution of people aged 65 and over in the states.

(c) Enter the data into your calculator's statistics list editor. Make a histogram using a window that matches your histogram from part (a). Copy the calculator histogram and mark the scales on your paper.

(d) Use the calculator's zoom feature to generate a histogram. Copy this histogram onto your paper and mark the scales.

(e) Store the data into the named list ELDER for later use.

1.13 DRP SCORES REVISITED Refer to Exercise 1.10 (page 17). Make a histogram of the DRP test scores for the sample of 44 children. Be sure to show your frequency table. Which do you prefer: the stemplot from Exercise 1.10 or the histogram that you just constructed? Why?

1.14 CEO SALARIES In 1993, *Forbes* magazine reported the age and salary of the chief executive officer (CEO) of each of the top 59 small businesses.[8] Here are the salary data, rounded to the nearest thousand dollars:

145	621	262	208	362	424	339	736	291	58	498	643	390	332
750	368	659	234	396	300	343	536	543	217	298	1103	406	254
862	204	206	250	21	298	350	800	726	370	536	291	808	543
149	350	242	198	213	296	317	482	155	802	200	282	573	388
250	396	572											

Construct a histogram for these data. Describe the shape, center, and spread of the distribution of CEO salaries. Are there any apparent outliers?

1.15 CHEST OUT, SOLDIER! In 1846, a published paper provided chest measurements (in inches) of 5738 Scottish militiamen. Table 1.6 displays the data in summary form.

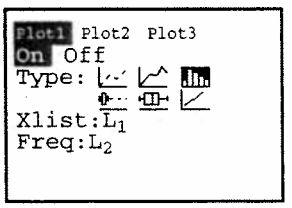
TABLE 1.6 Chest measurements (inches) of 5738 Scottish militiamen

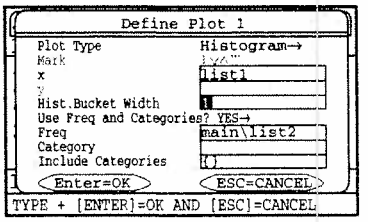
Chest size	Count	Chest size	Count
33	3	41	934
34	18	42	658
35	81	43	370
36	185	44	92
37	420	45	50
38	749	46	21
39	1073	47	4
40	1079	48	1

Source: Data and Story Library (DASL), http://lib.stat.cmu.edu/DASL/.

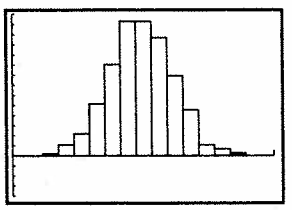
(a) You can use your graphing calculator to make a histogram of data presented in summary form like the chest measurements of Scottish militiamen.

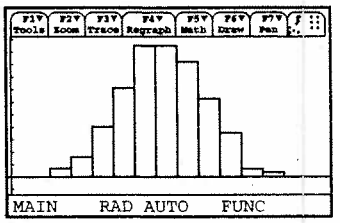
- Type the chest measurements into L_1/list1 and the corresponding counts into L_2/list2.
- Set up a statistics plot to make a histogram with x-values from L_1/list1 and y-values (bar heights) from L_2/list2.





- Adjust your viewing window settings as follows: xmin = 32, xmax = 49, xscl = 1, ymin = −300, ymax = 1100, yscl = 100. From now on, we will abbreviate in this form: $X[32,49]_1$ by $Y[-300,1100]_{100}$. Try using the calculator's built-in ZoomStat/ZoomData command. What happens?
- Graph.





(b) Describe the shape, center, and spread of the chest measurements distribution. Why might this information be useful?

More about shape

When you describe a distribution, concentrate on the main features. Look for major peaks, not for minor ups and downs in the bars of the histogram. Look for clear outliers, not just for the smallest and largest observations. Look for rough *symmetry* or clear *skewness*.

SYMMETRIC AND SKEWED DISTRIBUTIONS

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

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

In mathematics, symmetry means that the two sides of a figure like a histogram are exact mirror images of each other. Data are almost never exactly symmetric, so we are willing to call histograms like that in Exercise 1.15 approximately symmetric as an overall description. Here are more examples.

EXAMPLE 1.7 LIGHTNING FLASHES AND SHAKESPEARE

Figure 1.8 comes from a study of lightning storms in Colorado. It shows the distribution of the hour of the day during which the first lightning flash for that day occurred. The distribution has a single peak at noon and falls off on either side of this peak. The two sides of the histogram are roughly the same shape, so we call the distribution symmetric.

Figure 1.9 shows the distribution of lengths of words used in Shakespeare's plays.[9] This distribution also has a single peak but is skewed to the right. That is, there are many short words (3 and 4 letters) and few very long words (10, 11, or 12 letters), so that the right tail of the histogram extends out much farther than the left tail.

Notice that the vertical scale in Figure 1.9 is not the *count* of words but the *percent* of all of Shakespeare's words that have each length. A histogram of percents rather than counts is convenient when the counts are very large or when we want to compare several distributions. Different kinds of writing have different distributions of word lengths, but all are right-skewed because short words are common and very long words are rare.

The overall shape of a distribution is important information about a variable. Some types of data regularly produce distributions that are symmetric or skewed. For example, the sizes of living things of the same species (like lengths of cockroaches) tend to be symmetric. Data on incomes (whether of individuals, companies, or nations) are usually strongly skewed to the right. There are many moderate incomes, some large incomes, and a few very large incomes. Do remember that

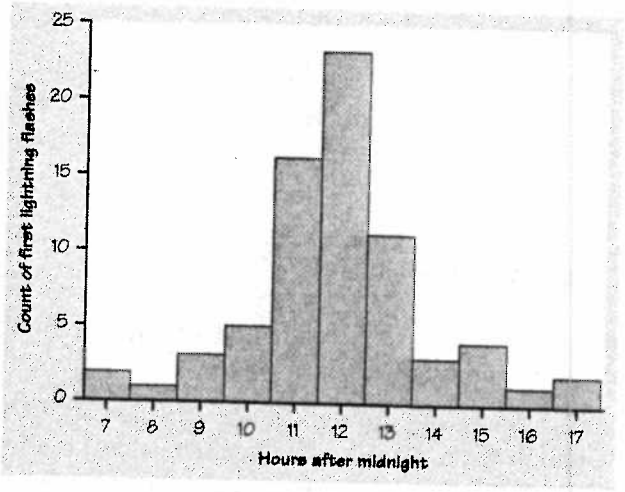


FIGURE 1.8 The distribution of the time of the first lightning flash each day at a site in Colorado, for Example 1.7.

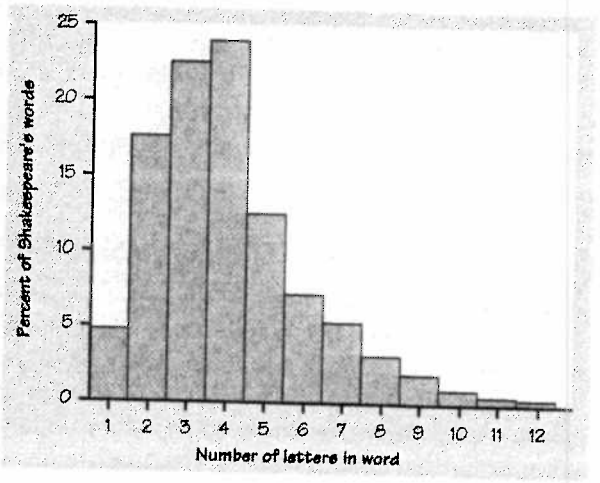


FIGURE 1.9 The distribution of lengths of words used in Shakespeare's plays, for Example 1.7.

many distributions have shapes that are neither symmetric nor skewed. Some data show other patterns. Scores on an exam, for example, may have a cluster near the top of the scale if many students did well. Or they may show two distinct peaks if a tough problem divided the class into those who did and didn't solve it. Use your eyes and describe what you see.

EXERCISES

1.16 **STOCK RETURNS** The total return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. Figure 1.10 is a histogram of the distribution of total returns for all 1528 stocks listed on the New York Stock Exchange in one year.[10] Like

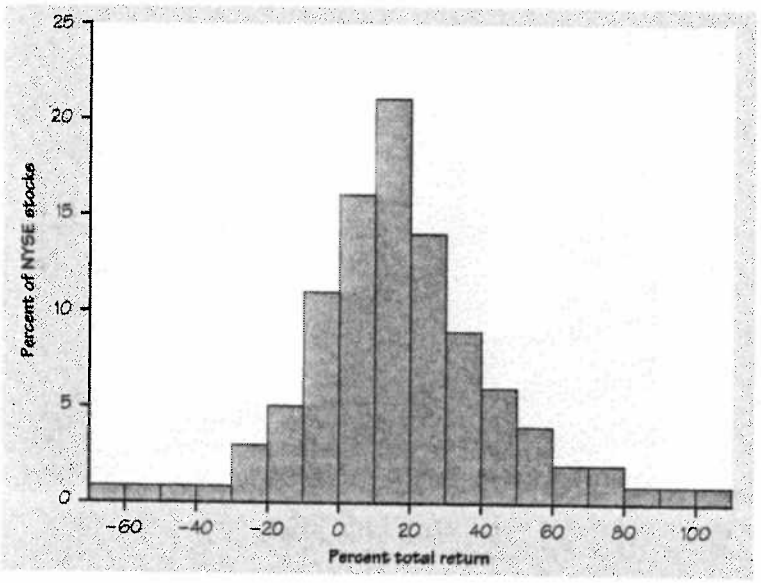


FIGURE 1.10 The distribution of percent total return for all New York Stock Exchange common stocks in one year.

Figure 1.9, it is a histogram of the percents in each class rather than a histogram of counts.

(a) Describe the overall shape of the distribution of total returns.

(b) What is the approximate center of this distribution? (For now, take the center to be the value with roughly half the stocks having lower returns and half having higher returns.)

(c) Approximately what were the smallest and largest total returns? (This describes the spread of the distribution.)

(d) A return less than zero means that an owner of the stock lost money. About what percent of all stocks lost money?

1.17 FREEZING IN GREENWICH, ENGLAND Figure 1.11 is a histogram of the number of days in the month of April on which the temperature fell below freezing at Greenwich, England.[11] The data cover a period of 65 years.

(a) Describe the shape, center, and spread of this distribution. Are there any outliers?

(b) In what percent of these 65 years did the temperature never fall below freezing in April?

1.18 How would you describe the center and spread of the distribution of first lightning flash times in Figure 1.8? Of the distribution of Shakespeare's word lengths in Figure 1.9?

Relative frequency, cumulative frequency, percentiles, and ogives

Sometimes we are interested in describing the relative position of an individual within a distribution. You may have received a standardized test score report that said you were in the 80th percentile. What does this mean? Put simply,

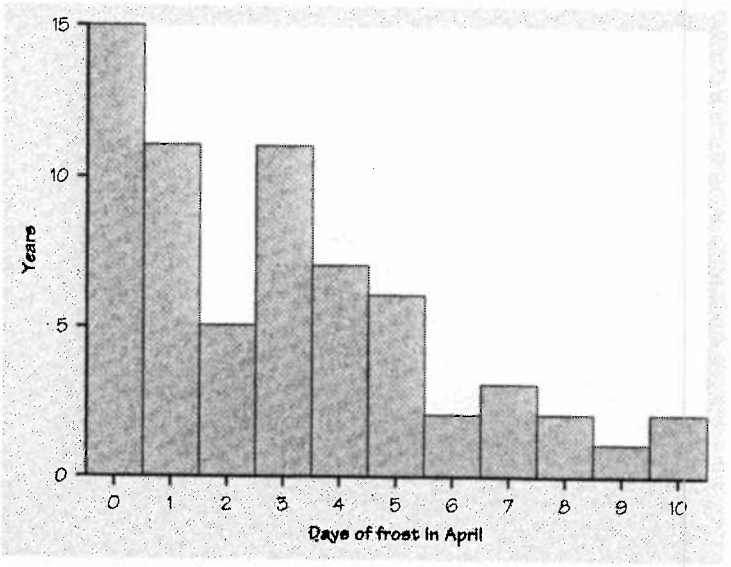


FIGURE 1.11 The distribution of the number of frost days during April at Greenwich, England, over a 65-year period, for Exercise 1.17.

80% of the people who took the test earned scores that were less than or equal to your score. The other 20% of students taking the test earned higher scores than you did.

PERCENTILE

The ***p*th percentile** of a distribution is the value such that p percent of the observations fall at or below it.

A histogram does a good job of displaying the distribution of values of a variable. But it tells us little about the relative standing of an individual observation. If we want this type of information, we should construct a **relative cumulative frequency graph,** often called an **ogive** (pronounced O-JIVE).

EXAMPLE 1.8 WAS BILL CLINTON A YOUNG PRESIDENT?

In Example 1.6, we made a histogram of the ages of U.S. presidents when they were inaugurated. Now we will examine where some specific presidents fall within the age distribution.

How to construct an ogive (relative cumulative frequency graph):

Step 1: Decide on class intervals and make a frequency table, just as in making a histogram. Add three columns to your frequency table: relative frequency, cumulative frequency, and relative cumulative frequency.

- To get the values in the *relative frequency* column, divide the count in each class interval by 43, the total number of presidents. Multiply by 100 to convert to a percentage.
- To fill in the *cumulative frequency* column, add the counts in the frequency column that fall in or below the current class interval.
- For the *relative cumulative frequency* column, divide the entries in the cumulative frequency column by 43, the total number of individuals.

Here is the frequency table from Example 1.6 with the relative frequency, cumulative frequency, and relative cumulative frequency columns added.

Class	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
40–44	2	$\frac{2}{43} = 0.047$, or 4.7%	2	$\frac{2}{43} = 0.047$, or 4.7%
45–49	6	$\frac{6}{43} = 0.140$, or 14.0%	8	$\frac{8}{43} = 0.186$, or 18.6%
50–54	13	$\frac{13}{43} = 0.302$, or 30.2%	21	$\frac{21}{43} = 0.488$, or 48.8%
55–59	12	$\frac{12}{43} = 0.279$, or 27.9%	33	$\frac{33}{43} = 0.767$, or 76.7%
60–64	7	$\frac{7}{43} = 0.163$, or 16.3%	40	$\frac{40}{43} = 0.930$, or 93.0%
65–69	3	$\frac{3}{43} = 0.070$, or 7.0%	43	$\frac{43}{43} = 1.000$, or 100%
TOTAL	43			

Step 2: Label and scale your axes and title your graph. Label the horizontal axis "Age at inauguration" and the vertical axis "Relative cumulative frequency." Scale the horizontal axis according to your choice of class intervals and the vertical axis from 0% to 100%.

Step 3: Plot a point corresponding to the relative cumulative frequency in each class interval at the *left endpoint* of the *next* class interval. For example, for the 40–44 interval, plot a point at a height of 4.7% above the age value of 45. This means that 4.7% of presidents were inaugurated before they were 45 years old. Begin your ogive with a point at a height of 0% at the left endpoint of the lowest class interval. Connect consecutive points with a line segment to form the ogive. The last point you plot should be at a height of 100%. Figure 1.12 shows the completed ogive.

How to locate an individual within the distribution:

What about Bill Clinton? He was age 46 when he took office. To find his relative standing, draw a vertical line up from his age (46) on the horizontal axis until it meets the ogive. Then draw a horizontal line from this point of intersection to the vertical axis. Based on Figure 1.13(a), we would estimate that Bill Clinton's age places him at the 10% *relative cumulative frequency* mark. That tells us that about 10% of all U.S. presidents were the same age as or younger than Bill Clinton when they were inaugurated. Put another way, President Clinton was younger than about 90% of all U.S. presidents based on his inauguration age. His age places him at the *10th percentile* of the distribution.

How to locate a value corresponding to a percentile:

- What inauguration age corresponds to the 60th percentile? To answer this question, draw a horizontal line across from the vertical axis at a height of 60% until it meets the ogive. From the point of intersection, draw a vertical line down to the horizontal axis.

In Figure 1.13(b), the value on the horizontal axis is about 57. So about 60% of all presidents were 57 years old or younger when they took office.

- Find the center of the distribution. Since we use the value that has half of the observations above it and half below it as our estimate of center, we simply need to find the 50th percentile of the distribution. Estimating as for the previous question, confirm that 55 is the center.

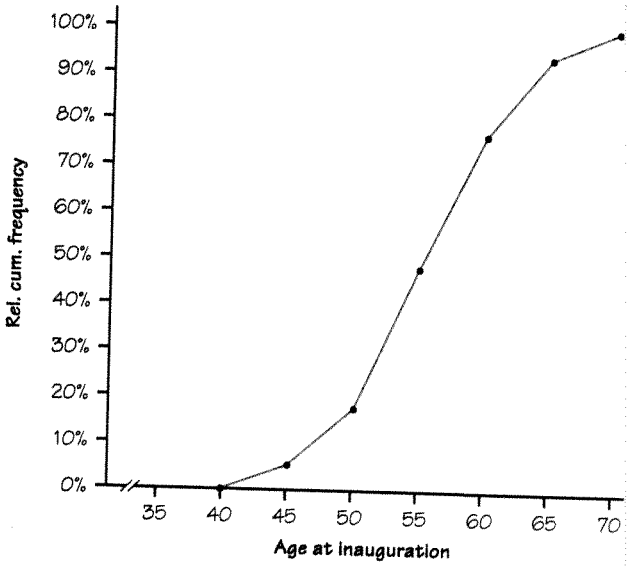


FIGURE 1.12 Relative cumulative frequency plot (ogive) for the ages of U.S. presidents at inauguration.

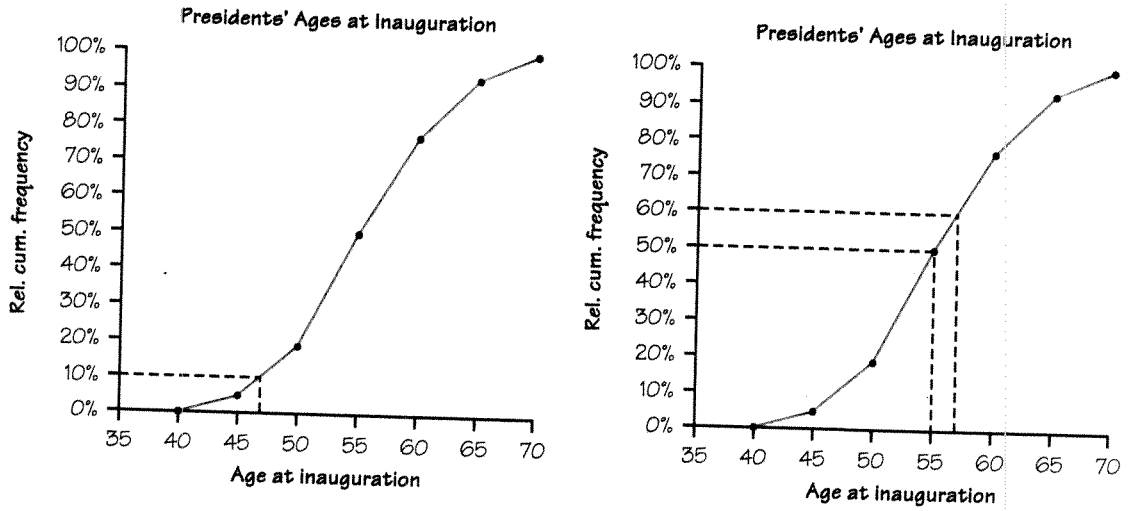


FIGURE 1.13 Ogives of presidents' ages at inauguration are used to (a) locate Bill Clinton within the distribution and (b) determine the 60th percentile and center of the distribution.

EXERCISES

1.19 OLDER FOLKS, II In Exercise 1.12 (page 22), you constructed a histogram of the percentage of people aged 65 or older in each state.

(a) Construct a relative cumulative frequency graph (ogive) for these data.

(b) Use your ogive from part (a) to answer the following questions:

- In what percentage of states was the percentage of "65 and older" less than 15%?
- What is the 40th percentile of this distribution, and what does it tell us?
- What percentile is associated with your state?

1.20 SHOPPING SPREE, II Figure 1.14 is an ogive of the amount spent by grocery shoppers in Exercise 1.11 (page 18).

(a) Estimate the center of this distribution. Explain your method.

(b) At what percentile would the shopper who spent $17.00 fall?

(c) Draw the histogram that corresponds to the ogive.

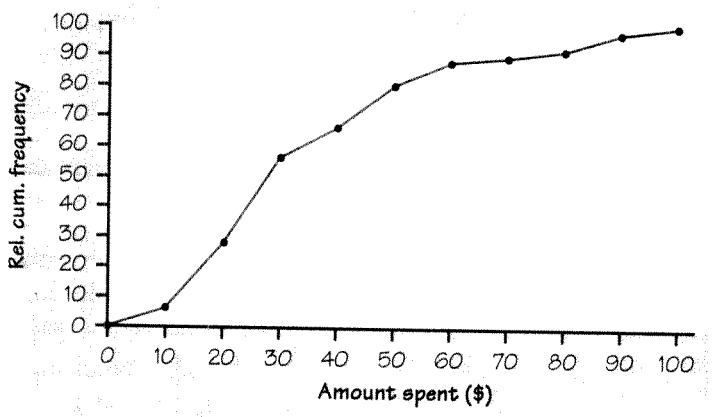


FIGURE 1.14 Amount spent by grocery shoppers in Exercise 1.11.

Time plots

Many variables are measured at intervals over time. We might, for example, measure the height of a growing child or the price of a stock at the end of each month. In these examples, our main interest is change over time. To display change over time, make a time plot.

TIME PLOT

A **time plot** of a variable plots each observation against the time at which it was measured. Always mark the time scale on the horizontal axis and the variable of interest on the vertical axis. If there are not too many points, connecting the points by lines helps show the pattern of changes over time.

trend

seasonal variation

When you examine a time plot, look once again for an overall pattern and for strong deviations from the pattern. One common overall pattern is a ***trend***, a long-term upward or downward movement over time. A pattern that repeats itself at regular time intervals is known as ***seasonal variation***. The next example illustrates both these patterns.

EXAMPLE 1.9 ORANGE PRICES MAKE ME SOUR!

Figure 1.15 is a time plot of the average price of fresh oranges over the period from January 1990 to January 2000. This information is collected each month as part of the government's reporting of retail prices. The vertical scale on the graph is the orange price index. This represents the price as a percentage of the average price of oranges in the years 1982 to 1984. The first value is 150 for January 1990, so at that time oranges cost about 150% of their 1982 to 1984 average price.

Figure 1.15 shows a clear *trend* of increasing price. In addition to this trend, we can see a strong *seasonal variation*, a regular rise and fall that occurs each year. Orange prices are usually highest in August or September, when the supply is lowest. Prices then fall in anticipation of the harvest and are lowest in January or February, when the harvest is complete and oranges are plentiful. The unusually large jump in orange prices in 1991 resulted from a freeze in Florida. Can you discover what happened in 1999?

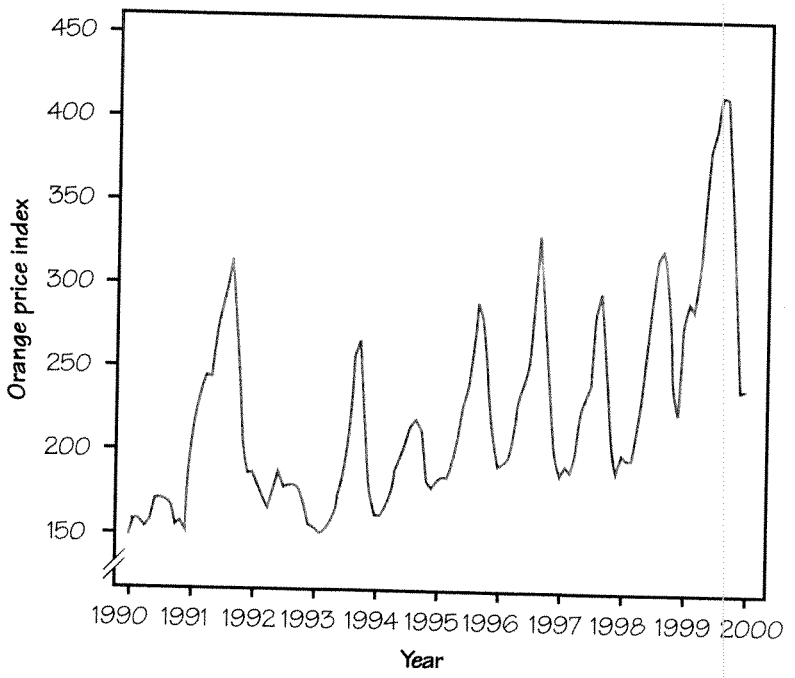


FIGURE 1.15 The price of fresh oranges, January 1990 to January 2000.

EXERCISES

1.21 CANCER DEATHS Here are data on the rate of deaths from cancer (deaths per 100,000 people) in the United States over the 50-year period from 1945 to 1995:

Year:	1945	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995
Deaths:	134.0	139.8	146.5	149.2	153.5	162.8	169.7	183.9	193.3	203.2	204.7

(a) Construct a time plot for these data. Describe what you see in a few sentences.

(b) Do these data suggest that we have made no progress in treating cancer? Explain.

1.22 CIVIL UNREST The years around 1970 brought unrest to many U.S. cities. Here are data on the number of civil disturbances in each three month period during the years 1968 to 1972:

Period		Count	Period		Count
1968	Jan.–Mar.	6	1970	July–Sept.	20
	Apr.–June	46		Oct.–Dec.	6
	July–Sept.	25	1971	Jan.–Mar.	12
	Oct.–Dec.	3		Apr.–June	21
1969	Jan.–Mar.	5		July–Sept.	5
	Apr.–June	27		Oct.–Dec.	1
	July–Sept.	19	1972	Jan.–Mar.	3
	Oct.–Dec.	6		Apr.–June	8
1970	Jan.–Mar.	26		July–Sept.	5
	Apr.–June	24		Oct.–Dec.	5

(a) Make a time plot of these counts. Connect the points in your plot by straight-line segments to make the pattern clearer.

(b) Describe the trend and the seasonal variation in this time series. Can you suggest an explanation for the seasonal variation in civil disorders?

SUMMARY

A data set contains information on a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, gender, or salary.

Exploratory data analysis uses graphs and numerical summaries to describe the variables in a data set and the relations among them.

Some variables are **categorical** and others are **quantitative.** A categorical variable places each individual into a category, like male or female. A quantitative variable has numerical values that measure some characteristic of each individual, like height in centimeters or annual salary in dollars.

The **distribution** of a variable describes what values the variable takes and how often it takes these values.

To describe a distribution, begin with a graph. Use **bar graphs** and **pie charts** to display categorical variables. **Dotplots, stemplots,** and **histograms** graph the distributions of quantitative variables. An **ogive** can help you determine relative standing within a quantitative distribution.

When examining any graph, look for an **overall pattern** and for notable **deviations** from the pattern.

The **center, spread,** and **shape** describe the overall pattern of a distribution. Some distributions have simple shapes, such as **symmetric** and **skewed.** Not all distributions have a simple overall shape, especially when there are few observations.

Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal **trends, seasonal variations,** or other changes over time.

SECTION 1.1 EXERCISES

1.23 GENDER EFFECTS IN VOTING Political party preference in the United States depends in part on the age, income, and gender of the voter. A political scientist selects a large sample of registered voters. For each voter, she records gender, age, household income, and whether they voted for the Democratic or for the Republican candidate in the last congressional election. Which of these variables are categorical and which are quantitative?

1.24 What type of graph or graphs would you plan to make in a study of each of the following issues?

(a) What makes of cars do students drive? How old are their cars?

(b) How many hours per week do students study? How does the number of study hours change during a semester?

(c) Which radio stations are most popular with students?

1.25 MURDER WEAPONS The 1999 *Statistical Abstract of the United States* reports FBI data on murders for 1997. In that year, 53.3% of all murders were committed with handguns, 14.5% with other firearms, 13.0% with knives, 6.3% with a part of the body (usually the hands or feet), and 4.6% with blunt objects. Make a graph to display these data. Do you need an "other methods" category?

1.26 WHAT'S A DOLLAR WORTH THESE DAYS? The buying power of a dollar changes over time. The Bureau of Labor Statistics measures the cost of a "market basket" of goods and services to compile its Consumer Price Index (CPI). If the CPI is 120, goods and services that cost $100 in the base period now cost $120. Here are the yearly average values of the CPI for the years between 1970 and 1999. The base period is the years 1982 to 1984.

Year	CPI	Year	CPI	Year	CPI	Year	CPI
1970	38.8	1978	65.2	1986	109.6	1994	148.2
1972	41.8	1980	82.4	1988	118.3	1996	156.9
1974	49.3	1982	96.5	1990	130.7	1998	163.0
1976	56.9	1984	103.9	1992	140.3	1999	166.6

(a) Construct a graph that shows how the CPI has changed over time.

(b) Check your graph by doing the plot on your calculator.

- Enter the years (the last two digits will suffice) into L_1/list1 and enter the CPI into L_2/list2.
- Then set up a statistics plot, choosing the plot type "xyline" (the second type on the TI-83). Use L_1/list1 as X and L_2/list2 as Y. In this graph, the data points are plotted and connected in order of appearance in L_1/list1 and L_2/list2.
- Use the zoom command to see the graph.

(c) What was the overall trend in prices during this period? Were there any years in which this trend was reversed?

(d) In what period during these decades were prices rising fastest? In what period were they rising slowest?

1.27 **THE STATISTICS OF WRITING STYLE** Numerical data can distinguish different types of writing, and sometimes even individual authors. Here are data on the percent of words of 1 to 15 letters used in articles in Popular Science magazine:[12]

Length:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Percent:	3.6	14.8	18.7	16.0	12.5	8.2	8.1	5.9	4.4	3.6	2.1	0.9	0.6	0.4	0.2

(a) Make a histogram of this distribution. Describe its shape, center, and spread.

(b) How does the distribution of lengths of words used in *Popular Science* compare with the similar distribution in Figure 1.9 (page 26) for Shakespeare's plays? Look in particular at short words (2, 3, and 4 letters) and very long words (more than 10 letters).

1.28 **DENSITY OF THE EARTH** In 1798 the English scientist Henry Cavendish measured the density of the earth by careful work with a torsion balance. The variable recorded was the density of the earth as a multiple of the density of water. Here are Cavendish's 29 measurements:[13]

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Present these measurements graphically in a stemplot. Discuss the shape, center, and spread of the distribution. Are there any outliers? What is your estimate of the density of the earth based on these measurements?

1.29 **DRIVE TIME** Professor Moore, who lives a few miles outside a college town, records the time he takes to drive to the college each morning. Here are the times (in minutes) for 42 consecutive weekdays, with the dates in order along the rows:

8.25	7.83	8.30	8.42	8.50	8.67	8.17	9.00	9.00	8.17	7.92
9.00	8.50	9.00	7.75	7.92	8.00	8.08	8.42	8.75	8.08	9.75
8.33	7.83	7.92	8.58	7.83	8.42	7.75	7.42	6.75	7.42	8.50
8.67	10.17	8.75	8.58	8.67	9.17	9.08	8.83	8.67		

(a) Make a histogram of these drive times. Is the distribution roughly symmetric, clearly skewed, or neither? Are there any clear outliers?

(b) Construct an ogive for Professor Moore's drive times.

(c) Use your ogive from (b) to estimate the center and 90th percentile for the distribution.

(d) Use your ogive to estimate the percentile corresponding to a drive time of 8.00 minutes.

1.30 **THE SPEED OF LIGHT** Light travels fast, but it is not transmitted instantaneously. Light takes over a second to reach us from the moon and over 10 billion years to reach us from the most distant objects observed so far in the expanding universe. Because radio and radar also travel at the speed of light, an accurate value for that speed is important in communicating with astronauts and orbiting satellites. An accurate value for the speed of light is also important to computer designers because electrical signals travel at light speed. The first reasonably accurate measurements of the speed of light were made over 100 years ago by A. A. Michelson and Simon Newcomb. Table 1.7 contains 66 measurements made by Newcomb between July and September 1882.

Newcomb measured the time in seconds that a light signal took to pass from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of about 7400 meters. Just as you can compute the speed of a car from the time required to drive a mile, Newcomb could compute the speed of light from the passage time. Newcomb's first measurement of the passage time of light was 0.000024828 second, or 24,828 nanoseconds. (There are 10^9 nanoseconds in a second.) The entries in Table 1.7 record only the deviation from 24,800 nanoseconds.

TABLE 1.7 Newcomb's measurements of the passage time of light

28	26	33	24	34	−44	27	16	40	−2	29	22	24	21
25	30	23	29	31	19	24	20	36	32	36	28	25	21
28	29	37	25	28	26	30	32	36	26	30	22	36	23
27	27	28	27	31	27	26	33	26	32	32	24	39	28
24	25	32	25	29	27	28	29	16	23				

Source: S. M. Stigler, "Do robust estimators work with real data?" *Annals of Statistics*, 5 (1977), pp. 1055–1078.

(a) Construct an appropriate graphical display for these data. Justify your choice of graph.

(b) Describe the distribution of Newcomb's speed of light measurements.

(c) Make a time plot of Newcomb's values. They are listed in order from left to right, starting with the top row.

(d) What does the time plot tell you that the display you made in part (a) does not?

Lesson: Sometimes you need to make more than one graphical display to uncover all of the important features of a distribution.

1.2 DESCRIBING DISTRIBUTIONS WITH NUMBERS

Who is baseball's greatest home run hitter? In the summer of 1998, Mark McGwire and Sammy Sosa captured the public's imagination with their pursuit of baseball's single-season home run record (held by Roger Maris). McGwire eventually set a new standard with 70 home runs. Barry Bonds broke Mark McGwire's record when he hit 73 home runs in the 2001 season. How does this accomplishment fit Bonds's career? Here are Bonds's home run counts for the years 1986 (his rookie year) to 2001 (the year he broke McGwire's record):

1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
16	25	24	19	33	25	34	46	37	33	42	40	37	34	49	73

The stemplot in Figure 1.16 shows us the *shape, center,* and *spread* of these data. The distribution is roughly symmetric with a single peak and a possible high outlier. The center is about 34 home runs, and the spread runs from 16 to the record 73. Shape, center, and spread provide a good description of the overall pattern of any distribution for a quantitative variable. Now we will learn specific ways to use numbers to measure the center and spread of a distribution.

```
1 | 6 9
2 | 4
2 | 5 5
3 | 3 3 4 4
3 | 7 7
4 | 0 2
4 | 6 9
5 |
5 |
6 |
6 |
7 | 3
```

FIGURE 1.16 Number of home runs hit by Barry Bonds in each of his 16 major league seasons.

Measuring center: the mean

A description of a distribution almost always includes a measure of its center or average. The most common measure of center is the ordinary arithmetic average, or *mean*.

THE MEAN $\bar{x}$

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, \ldots, x_n$, their mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or in more compact notation,

$$\bar{x} = \frac{1}{n}\sum x_i$$

The Σ (capital Greek sigma) in the formula for the mean is short for "add them all up." The subscripts on the observations x_i are just a way of keeping the n observations distinct. They do not necessarily indicate order or any other special facts about the data. The bar over the x indicates the mean of all the x-values. Pronounce the mean $\bar{x}$ as "x-bar." This notation is very common. When writers who are discussing data use $\bar{x}$ or $\bar{y}$, they are talking about a mean.

EXAMPLE 1.10 BARRY BONDS VERSUS HANK AARON

The mean number of home runs Barry Bonds hit in his first 16 major league seasons is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{16 + 25 + \cdots + 73}{16} = \frac{567}{16} = 35.4375$$

We might compare Bonds to Hank Aaron, the all-time home run leader. Here are the numbers of home runs hit by Hank Aaron through his last year with Atlanta:

13	27	26	44	30	39	40	34	45	44	24
32	44	39	29	44	38	47	34	40	20	

Aaron's mean number of home runs hit in a year is

$$\bar{x} = \frac{1}{21}(13 + 27 + \cdots + 20) = \frac{733}{21} = 34.9$$

Barry Bonds's exceptional performance in 2001 stands out from his home run production in the previous 15 seasons. Use your calculator to check that his mean home run production in his first 15 seasons is $\bar{x} = 32.93$. One outstanding season increased Bonds's mean home run count by 2.5 home runs per year.

Example 1.10 illustrates an important fact about the mean as a measure of center: it is sensitive to the influence of a few extreme observations. These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a ***resistant measure*** of center.

resistant measure

Measuring center: the median

In Section 1.1, we used the midpoint of a distribution as an informal measure of center. The *median* is the formal version of the midpoint, with a specific rule for calculation.

THE MEDIAN *M*

The **median** M is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list.

Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of observations in order is very tedious, however, so that finding the median by hand for larger sets of data is unpleasant. You will need computer software or a graphing calculator to automate finding the median.

EXAMPLE 1.11 FINDING MEDIANS

To find the median number of home runs Barry Bonds hit in his first 16 seasons, first arrange the data in increasing order:

16	19	24	25	25	33	33	**34**	**34**	37	37	40	42	46	49	73

The count of observations $n = 16$ is even. There is no center observation, but there is a center pair. These are the two bold 34s in the list, which have 7 observations to their left in the list and 7 to their right. The median is midway between these two observations. Because both of the middle pair are 34, $M = 34$.

How much does the apparent outlier affect the median? Drop the 73 from the list and find the median for the remaining $n = 15$ years. It is the 8th observation in the edited list, $M = 34$.

How does Bonds's median compare with Hank Aaron's? Here, arranged in increasing order, are Aaron's home run counts:

13	20	24	26	27	29	30
32	34	34	**38**	39	39	40
40	44	44	44	44	45	47

The number of observations is odd, so there is one center observation. This is the median. It is the bold 38, which has 10 observations to its left in the list and 10 observations to its right. Bonds now holds the single-season record, but he has hit fewer home runs in a typical season than Aaron. Barry Bonds also has a long way to go to catch Aaron's career total of 733 home runs.

Comparing the mean and the median

Examples 1.10 and 1.11 illustrate an important difference between the mean and the median. The one high value pulls Bonds's mean home run count up from 32.93 to 35.4375. The median is not affected at all. The median, unlike the mean, is *resistant*. If Bonds's record 73 had been 703, his median would not change at all. The 703 just counts as one observation above the center, no matter how far above the center it lies. The mean uses the actual value of each observation and so will chase a single large observation upward.

The mean and median of a symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is farther out in the long tail than is the median. For example, the distribution of house prices is strongly skewed to the right. There are many moderately priced houses and a few very expensive mansions. The few expensive houses pull the mean up but do not affect the median. The mean price of new houses sold in 1997 was $176,000, but the median price for these same houses was only $146,000. Reports about house prices, incomes, and other strongly skewed distributions usually give the median ("midpoint") rather than the mean ("arithmetic average"). However, if you are a tax assessor interested in the total value of houses in your area, use the mean. The total value is the mean times the number of houses; it has no connection with the median. The mean and median measure center in different ways, and both are useful.

EXERCISES

1.31 Joey's first 14 quiz grades in a marking period were

86	84	91	75	78	80	74	87	76	96	82	90	98	93

(a) Use the formula to calculate the mean. Check using "one-variable statistics" on your calculator.

(b) Suppose Joey has an unexcused absence for the fifteenth quiz and he receives a score of zero. Determine his final quiz average. What property of the mean does this situation illustrate? Write a sentence about the effect of the zero on Joey's quiz average that mentions this property.

(c) What kind of plot would best show Joey's distribution of grades? Assume an 8-point grading scale (A: 93 to 100, B: 85 to 92, etc.). Make an appropriate plot, and be prepared to justify your choice.

1.32 **SSHA SCORES** The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates college students' motivation, study habits, and attitudes toward school. A private college gives the SSHA to a sample of 18 of its incoming first-year women students. Their scores are

154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

(a) Make a stemplot of these data. The overall shape of the distribution is irregular, as often happens when only a few observations are available. Are there any potential outliers? About where is the center of the distribution (the score with half the scores above it and half below)? What is the spread of the scores (ignoring any outliers)?

(b) Find the mean score from the formula for the mean. Then enter the data into your calculator. You can find the mean from the home screen as follows:

TI-83	TI-89
• Press 2nd STAT (LIST) ▶▶ (MATH).	• Press CATALOG then 5 (M).
• Choose 3:mean(, enter list name, press ENTER.	• Choose mean(, type list name, press ENTER.

(c) Find the median of these scores. Which is larger: the median or the mean? Explain why.

1.33 Suppose a major league baseball team's mean yearly salary for a player is $1.2 million, and that the team has 25 players on its active roster. What is the team's annual payroll for players? If you knew only the median salary, would you be able to answer the question? Why or why not?

1.34 Last year a small accounting firm paid each of its five clerks $22,000, two junior accountants $50,000 each, and the firm's owner $270,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary? Write a sentence to describe how an unethical recruiter could use statistics to mislead prospective employees.

1.35 **U.S. INCOMES** The distribution of individual incomes in the United States is strongly skewed to the right. In 1997, the mean and median incomes of the top 1% of Americans were $330,000 and $675,000. Which of these numbers is the mean and which is the median? Explain your reasoning.

SOLUTIONS TO ODD NUMBERED EXERCISES

Chapter 1

1.1 (a) Cars. (b) Vehicle type and transmission type are categorical; number of cylinders, city mpg, and highway mpg are quantitative.

1.3 A few possibilities are: length (no. of pages), weight (ounces), thickness (cm).

1.5 (b) No. These percentages cannot be combined to form a whole (sum exceeds 100%).

1.7 Shape: skewed right; Center: 1; Spread: 0 to 39; Outliers: 26, 28, 39.

1.9 (a) Stems: thousands of dollars; Leaves: hundreds of dollars; data are rounded to nearest \$100. (b) Shape: skewed right; Center: about \$4500; Spread: about \$1300 to \$19,300; Outliers: maybe \$18,200 and \$19,300.

1.11 (a) The stemplot is skewed right with peak on the "2" stem. (b) This stemplot shows more distinct clusters and gaps. (c) Shape: skewed right; Center: about \$28; Spread: about \$3 to \$93. Over half the shoppers spent \$28 or less, but a few spent over \$80.

1.13 The histogram has a fairly indistinct shape and is centered at 35. There are no apparent outliers. The stemplot has the advantage of showing actual data values.

1.15 (a) The calculator's zoom feature may choose an unusual x-scale. (b) Shape: roughly symmetric; Center: 40 inches; Spread: 33 to 48 inches; might be useful in designing uniforms.

1.17 (a) Shape: skewed right; Center: 3 days; Spread: 0 to 10 days; Outliers: none. (b) 23.

1.19 (a) Answers will vary depending on choice of class intervals. (b) 90; 12.4%, meaning that 40% of states have 12.4% or less of their populations aged 65 or older.

1.21 (a) The timeplot shows an increasing trend in the cancer death rate from 1945 to 1995. (b) No; our ability to diagnose cancer has improved and elderly people make up a much higher proportion of the population now.

1.23 Categorical: Gender, vote; Quantitative: Age, household income.

1.25 You can make a bar graph or a pie chart. You will need an "other methods" category for the pie chart.

1.27 (a) Shape: skewed right; Center: 4 letters; Spread: 1 to 15 letters. (b) There are more 2-, 3-, and 4-letter words in Shakespeare's plays and more very long words in *Popular Science* articles.

1.29 (a) Roughly symmetric; no. (b) Answers will vary. (c) 8.42 min.; about 9 min. (d) About the 28th percentile.

1.31 (a) 85. (b) 79.33; the mean is not resistant to outliers. (c) A histogram.

1.33 \$30 million; No.

1.35 \$675,000 is the mean and \$330,000 is the median. A few extremely high salaries will pull the mean upward.

1.37 (a) About the same since the graph is symmetric. (b) min = 42, $Q_1 = 51$, $M = 55$, $Q_3 = 58$, max = 69 (c) 7 years (d) The boxplot is roughly symmetric. (e) Yes; Ronald Reagan.

1.39 (a) \$1735 (b) The boxplot is skewed right. The shoppers who spent \$85.76, \$86.37, and \$93.34 are outliers.

1.41 (a) $s = 15.526$ (b) $\bar{x} = 22.3$ and $s = 10.149$; No.

1.43 (a) All four numbers the same. (b) 0, 0, 10, 10 (c) For (a).