

H. COMPARING DISTRIBUTIONS

1. Use side-by-side bar graphs to compare distributions of categorical data.
2. Make back-to-back stemplots and side-by-side boxplots to compare distributions of quantitative variables.
3. Write narrative comparisons of the shape, center, spread, and outliers for two or more quantitative distributions.

CHAPTER 1 REVIEW EXERCISES

1.59 Each year *Fortune* magazine lists the top 500 companies in the United States, ranked according to their total annual sales in dollars. Describe three other variables that could reasonably be used to measure the “size” of a company.

1.60 **ATHLETES' SALARIES** Here is a small part of a data set that describes major league baseball players as of opening day of the 1998 season:

Player	Team	Position	Age	Salary
:				
Perez, Eduardo	Reds	First base	28	300
Perez, Neifi	Rockies	Shortstop	23	210
Pettitte, Andy	Yankees	Pitcher	25	3750
Piazza, Mike	Dodgers	Catcher	29	8000
:				

- (a) What individuals does this data set describe?
- (b) In addition to the player's name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?
- (c) Based on the data in the table, what do you think are the units of measurement for each of the quantitative variables?

1.61 **HOW YOUNG PEOPLE DIE** The number of deaths among persons aged 15 to 24 years in the United States in 1997 due to the seven leading causes of death for this age group were accidents, 12,958; homicide, 5793; suicide, 4146; cancer, 1583; heart disease, 1013; congenital defects, 383; AIDS, 276.¹⁷

- (a) Make a bar graph to display these data.
- (b) What additional information do you need to make a pie chart?

1.62 **NEVER ON SUNDAY?** The Canadian Province of Ontario carries out statistical studies of the working of Canada's national health care system in the province. The bar graphs in Figure 1.24 come from a study of admissions and discharges from community hospitals in Ontario.¹⁸ They show the number of heart attack patients admitted and discharged on each day of the week during a 2-year period.

- (a) Explain why you expect the number of patients admitted with heart attacks to be roughly the same for all days of the week. Do the data show that this is true?
- (b) Describe how the distribution of the day on which patients are discharged from the hospital differs from that of the day on which they are admitted. What do you think explains the difference?

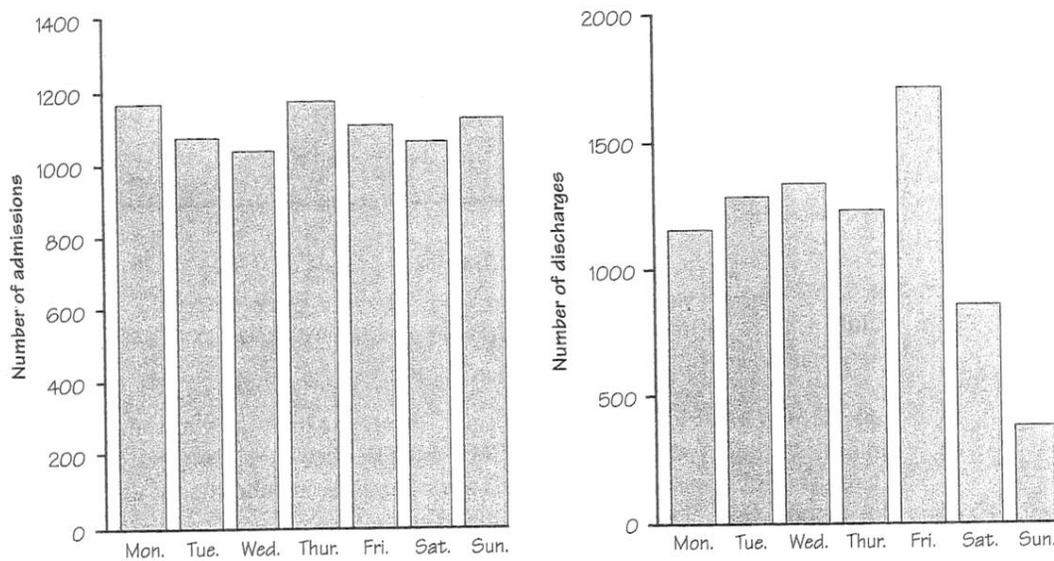


FIGURE 1.24 Bar graphs of the number of heart attack victims admitted and discharged on each day of the week by hospitals in Ontario, Canada.

1.63 PRESIDENTIAL ELECTIONS Here are the percents of the popular vote won by the successful candidate in each of the presidential elections from 1948 to 2000:

Year:	1948	1952	1956	1960	1964	1968	1972	1976	1980	1984	1988	1992	1996	2000
Percent:	49.6	55.1	57.4	49.7	61.1	43.4	60.7	50.1	50.7	58.8	53.9	43.2	49.2	47.9

- (a) Make a stemplot of the winners' percents. (Round to whole numbers and use split stems.)
- (b) What is the median percent of the vote won by the successful candidate in presidential elections? (Work with the unrounded data.)
- (c) Call an election a landslide if the winner's percent falls at or above the third quartile. Find the third quartile. Which elections were landslides?

1.64 HURRICANES The histogram in Figure 1.25 (next page) shows the number of hurricanes reaching the east coast of the United States each year over a 70-year period.¹⁹ Give a brief description of the overall shape of this distribution. About where does the center of the distribution lie?

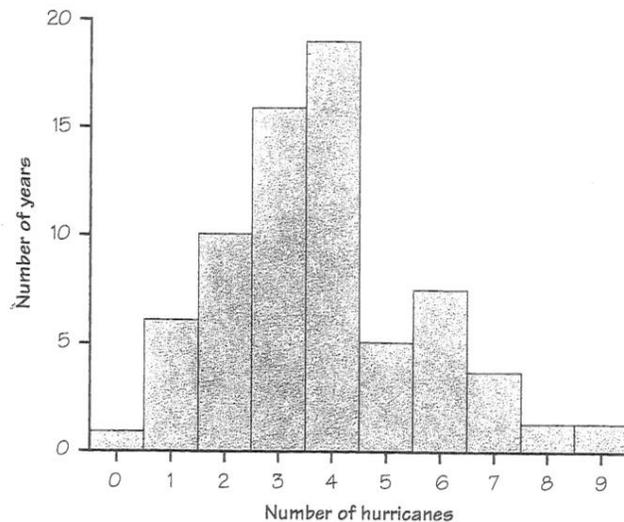


FIGURE 1.25 The distribution of the annual number of hurricanes on the U.S. east coast over a 70-year period, for Exercise 1.64.

1.65 DO SUVs WASTE GAS? Table 1.3 (page 17) gives the highway fuel consumption (in miles per gallon) for 32 model year 2000 midsize cars. We constructed a dotplot for these data in Exercise 1.8. Table 1.14 shows the highway mileages for 26 four-wheel-drive model year 2000 sport utility vehicles.

- Give a graphical and numerical description of highway fuel consumption for SUVs. What are the main features of the distribution?
- Make boxplots to compare the highway fuel consumption of midsize cars and SUVs. What are the most important differences between the two distributions?

TABLE 1.14 Highway gas mileages for model year 2000 four-wheel-drive SUVs

Model	MPG	Model	MPG
BMW X5	17	Kia Sportage	22
Chevrolet Blazer	20	Land Rover	17
Chevrolet Tahoe	18	Lexus LX470	16
Dodge Durango	18	Lincoln Navigator	17
Ford Expedition	18	Mazda MPV	19
Ford Explorer	20	Mercedes-Benz ML320	20
Honda Passport	20	Mitsubishi Montero	20
Infinity QX4	18	Nissan Pathfinder	19
Isuzu Amigo	19	Nissan Xterra	19
Isuzu Trooper	19	Subaru Forester	27
Jeep Cherokee	20	Suzuki Grand Vitara	20
Jeep Grand Cherokee	18	Toyota RAV4	26
Jeep Wrangler	19	Toyota 4Runner	21

1.66 DR. DATA RETURNS! Dr. Data asked her students how much time they spent using a computer during the previous week. Figure 1.26 is an ogive of her students' responses.

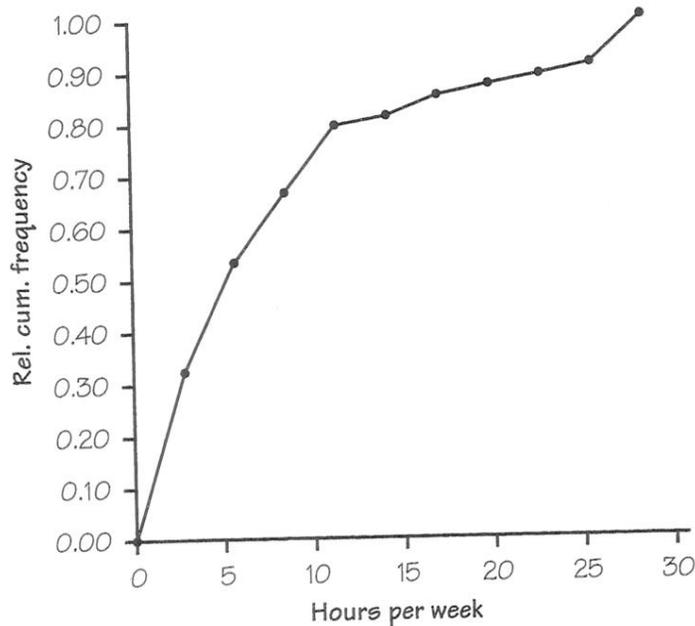


FIGURE 1.26 Ogive of weekly computer use by Dr. Data's statistics students.

- Construct a relative frequency table based on the ogive. Then make a histogram.
- Estimate the median, Q_1 , and Q_3 from the ogive. Then make a boxplot. Are there any outliers?
- At what percentile does a student who used her computer for 10 hours last week fall?

1.67 WAL-MART STOCK The rate of return on a stock is its change in price plus any dividends paid. Rate of return is usually measured in percent of the starting value. We have data on the monthly rates of return for the stock of Wal-Mart stores for the years 1973 to 1991, the first 19 years Wal-Mart was listed on the New York Stock Exchange. There are 228 observations.

Figure 1.27 (next page) displays output from statistical software that describes the distribution of these data. The stems in the stemplot are the tens digits of the percent returns. The leaves are the ones digits. The stemplot uses split stems to give a better display. The software gives high and low outliers separately from the stemplot rather than spreading out the stemplot to include them.

- Give the five-number summary for monthly returns on Wal-Mart stock.
- Describe in words the main features of the distribution.
- If you had \$1000 worth of Wal-Mart stock at the beginning of the best month during these 19 years, how much would your stock be worth at the end of the month? If you had \$1000 worth of stock at the beginning of the worst month, how much would your stock be worth at the end of the month?
- Find the interquartile range (IQR) for the Wal-Mart data. Are there any outliers according to the $1.5 \times \text{IQR}$ criterion? Does it appear to you that the software uses this criterion in choosing which observations to report separately as outliers?

```

Mean = 3.064
Standard deviation = 11.49

N = 228   Median = 3.4691
Quartiles = -2.950258, 8.4511

Decimal point is 1 place to the right of the colon

Low:  -34.04255  -31.25000  -27.06271  -26.61290

-1 : 985
-1 : 444443322222110000
-0 : 99998877766666665555
-0 : 44444443333332222222111111100
 0 : 0000111111111112222233333344444444
 0 : 55555555555555556666666666777777888888888899999
 1 : 000000001111111122233334444
 1 : 55566667889
 2 : 011334

High:  32.01923  41.80531  42.05607  57.89474  58.67769

```

FIGURE 1.27 Output from software describing the distribution of monthly returns from Wal-Mart stock.

1.68 A study of the size of jury awards in civil cases (such as injury, product liability and medical malpractice) in Chicago showed that the median award was about \$8000. But the mean award was about \$69,000. Explain how this great difference between the two measures of center can occur.

1.69 You want to measure the average speed of vehicles on the interstate highway on which you are driving. You adjust your speed until the number of vehicles passing you equals the number you are passing. Have you found the mean speed or the median speed of vehicles on the highway?

TABLE 1.15 Data on education in the United States for Exercises 1.70 to 1.73

State	Region	Population (1000)	SAT Verbal	SAT Math	Percent taking	Percent no HS diploma	Teachers' pay (\$1000)
AL	ESC	4,447	561	555	9	33.1	32.8
AK	PAC	627	516	514	50	13.4	51.7
AZ	MTN	5,131	524	525	34	21.3	34.4
AR	WSC	2,673	563	556	6	33.7	30.6
CA	PAC	33,871	497	514	49	23.8	43.7

TABLE 1.15 Data on education in the United States, for Exercises 1.70 to 1.73
(continued)

State	Region	Population (1000)	SAT Verbal	SAT Math	Percent taking	Percent no HS diploma	Teachers' pay (\$1000)
CO	MTN	4,301	536	540	32	15.6	37.1
CT	NE	3,406	510	509	80	20.8	50.7
DE	SA	784	503	497	67	22.5	42.4
DC	SA	572	494	478	77	26.9	46.4
FL	SA	15,982	499	498	53	25.6	34.5
GA	SA	8,186	487	482	63	29.1	37.4
HI	PAC	1,212	482	513	52	19.9	38.4
ID	MTN	1,294	542	540	16	20.3	32.8
IL	ENC	12,419	569	585	12	23.8	43.9
IN	ENC	6,080	496	498	60	24.4	39.7
IA	WNC	2,926	594	598	5	19.9	34.0
KS	WNC	2,688	578	576	9	18.7	36.8
KY	ESC	4,042	547	547	12	35.4	34.5
LA	WSC	4,469	561	558	8	31.7	29.7
ME	NE	1,275	507	503	68	21.2	34.3
MD	SA	5,296	507	507	65	21.6	41.7
MA	NE	6,349	511	511	78	20.0	43.9
MI	ENC	9,938	557	565	11	23.2	49.3
MN	WNC	4,919	586	598	9	17.6	39.1
MS	ESC	2,845	563	548	4	35.7	29.5
MO	WNC	5,595	572	572	8	26.1	34.0
MT	MTN	902	545	546	21	19.0	30.6
NE	WNC	1,711	568	571	8	18.2	32.7
NV	MTN	1,998	512	517	34	21.2	37.1
NH	NE	1,236	520	518	72	17.8	36.6
NJ	MA	8,414	498	510	80	23.3	50.4
NM	MTN	1,819	549	542	12	24.9	30.2
NY	MA	18,976	495	502	76	25.2	49.0
NC	SA	8,049	493	493	61	30.0	33.3
ND	WNC	642	594	605	5	23.3	28.2
OH	ENC	11,353	534	568	25	24.3	39.0
OK	WSC	3,451	567	560	8	25.4	30.6
OR	PAC	3,421	525	525	53	18.5	42.2
PA	MA	12,281	498	495	70	25.3	47.7
RI	NE	1,048	504	499	70	28.0	44.3
SC	SA	4,012	479	475	61	31.7	33.6
SD	WNC	755	585	588	4	22.9	27.3
TN	ESC	5,689	559	553	13	32.9	35.3
TX	WSC	20,852	494	499	50	27.9	33.6
UT	MTN	2,233	570	568	5	14.9	33.0
VT	NE	609	514	506	70	19.2	36.3
VA	SA	7,079	508	499	65	24.8	36.7
WA	PAC	5,894	525	526	52	16.2	38.8
WV	SA	1,808	527	512	18	34.0	33.4
WI	ENC	5,364	584	595	7	21.4	39.9
WY	MTN	494	546	551	10	17.0	32.0

Source: U.S. Census Bureau Web site, <http://www.census.gov>, 2001.

Table 1.15 presents data about the individual states that relate to education. Study of a data set with many variables begins by examining each variable by itself. Exercises 1.70 to 1.73 concern the data in Table 1.15.

1.70 POPULATION OF THE STATES Make a graphical display of the population of the states. Briefly describe the shape, center, and spread of the distribution of population. Explain why the shape of the distribution is not surprising. Are there any states that you consider outliers?

1.71 HOW MANY STUDENTS TAKE THE SAT? Make a stemplot of the distribution of the percent of high school seniors who take the SAT in the various states. Briefly describe the overall shape of the distribution. Find the midpoint of the data and mark this value on your stemplot. Explain why describing the center is not very useful for a distribution with this shape.

1.72 HOW MUCH ARE TEACHERS PAID? Make a graph to display the distribution of average teachers' salaries for the states. Is there a clear overall pattern? Are there any outliers or other notable deviations from the pattern?

1.73 PEOPLE WITHOUT HIGH SCHOOL EDUCATIONS The "Percent no HS" column gives the percent of the adult population in each state who did not graduate from high school. We want to compare the percents of people without a high school education in the northeastern and the southern states. Take the northeastern states to be those in the MA (Mid-Atlantic) and NE (New England) regions. The southern states are those in the SA (South Atlantic) and ESC (East South Central) regions. Leave out the District of Columbia, which is a city rather than a state.

(a) List the percents without high school for the northeastern and for the southern states from Table 1.15. These are the two data sets we want to compare.

(b) Make numerical summaries and graphs to compare the two distributions. Write a brief statement of what you find.

NOTES AND DATA SOURCES

1. Data from *Beverage Digest*, February 18, 2000.
2. Seat-belt data from the National Highway and Traffic Safety Administration, *NOPUS Survey*, 1998.
3. Data from the 1997 *Statistical Abstract of the United States*.
4. Data on accidental deaths from the Centers for Disease Control Web site, www.cdc.gov.
5. Data from the *Los Angeles Times*, February 16, 2001.
6. Based on experiments performed by G. T. Lloyd and E. H. Ramshaw of the CSIRO Division of Food Research, Victoria, Australia, 1982–83.
7. Maribeth Cassidy Schmitt, from her Ph.D. dissertation, "The effects of an elaborated directed reading activity on the metacomprehension skills of third graders," Purdue University, 1987.
8. Data from "America's best small companies," *Forbes*, November 8, 1993.
9. The Shakespeare data appear in C. B. Williams, *Style and Vocabulary: Numerological Studies*, Griffin, London, 1970.

