- For categorical variables, a ***contingency table*** shows the number of observations at the combinations of possible outcomes for the two variables.
- For quantitative variables, a ***scatterplot*** graphs the observations, showing a point for each observation. The response variable is plotted on the y-axis and the explanatory variable is plotted on the x-axis.
- For quantitative variables, the ***correlation*** describes the strength of straight-line association. It falls between -1 and $+1$ and indicates whether the response variable tends to increase (positive correlation) or decrease (negative correlation) as the explanatory variable increases.
- A ***regression analysis*** provides a straight-line formula for predicting the value of the response variable using the explanatory variable. We study correlation and regression in detail in Chapter 9.

PROBLEMS

Practicing the Basics

3.1. Table 3.10 shows the number (in millions) of the foreign-born population of the United States in 2004, by place of birth.

(a) Construct a relative frequency distribution.
(b) Sketch the data in a bar graph.
(c) Is "Place of birth" quantitative or categorical?
(d) Use whichever of the following measures is relevant for these data: mean, median, mode.

TABLE 3.10

Place of Birth	Number
Europe	4.7
Asia	8.7
Caribbean	3.3
Central America	12.9
South America	2.1
Other	2.6
Total	34.3

Source: Statistical Abstract of the United States, 2006.

3.2. According to www.adherents.com, in 2006 the number of followers of the world's five largest religions were 2.1 billion for Christianity, 1.3 billion for Islam, 0.9 billion for Hinduism, 0.4 billion for Confucianism, and 0.4 billion for Buddhism.

(a) Construct a relative frequency distribution.
(b) Sketch a bar graph.
(c) Can you find a mean, median, or mode for these data? If so, do so and interpret.

3.3. A teacher shows her class the scores on the midterm exam in the stem-and-leaf plot:

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

(a) Identify the number of students and the minimum and maximum scores.
(b) Sketch a histogram with four intervals.

3.4. According to the *2005 American Community Survey*, in 2005 the United States had 30.1 million households with one person, 37.0 million with two persons, 17.8 million with three persons, 15.3 million with four persons, and 10.9 million with five or more persons.

(a) Construct a relative frequency distribution.
(b) Sketch a histogram. What is its shape?
(c) Report and interpret the (i) median, (ii) mode of household size.

3.5. Copy the "2005 statewide crime" data file from the text Web site (www.stat.ufl.edu/~aa/social/data.html). Use the variable, murder rate (per 100,000 population). In this exercise, do not use the observation for D.C. Using software,

(a) Construct a relative frequency distribution.
(b) Construct a histogram. How would you describe the shape of the distribution?
(c) Construct a stem-and-leaf plot. How does this plot compare to the histogram in (b)?

3.6. The OECD (Organization for Economic Cooperation and Development) consists of advanced, industrialized countries that accept the principles of representative democracy and a free market economy. Table 3.11 shows UN data for OECD nations on several variables: gross domestic product (GDP, per capita in U.S. dollars), percent unemployed, a measure of inequality based on comparing wealth of the richest 10% to the poorest 10%, public expenditure on health (as a percentage of the GDP), the number of physicians per 100,000 people, carbon dioxide emissions (per capita, in metric tons), the percentage of seats in parliament held by women, and female economic activity as

TABLE 3.11: UN Data for OECD Nations, Available as "OECD data" File at Text Web Site

Nation	GDP	Unemp.	Inequal.	Health	Physicians	C02	Women Parl.	Fem. Econ.
Australia	30,331	5.1	12.5	6.4	247	18	28.3	79
Austria	32,276	5.8	6.9	5.1	338	8.6	32.2	75
Belgium	31,096	8.4	8.2	6.3	449	8.3	35.7	72
Canada	31,263	6.8	9.4	6.9	214	17.9	24.3	83
Denmark	31,914	4.9	8.1	7.5	293	10.1	36.9	84
Finland	29,951	8.6	5.6	5.7	316	13	37.5	86
France	29,300	10.0	9.1	7.7	337	6.2	13.9	79
Germany	28,303	9.3	6.9	8.7	337	9.8	30.5	76
Greece	22,205	10.6	10.2	5.1	438	8.7	13	66
Iceland	33,051	2.5	..	8.8	362	7.6	33.3	87
Ireland	38,827	4.3	9.4	5.8	279	10.3	14.2	72
Italy	28,180	7.7	11.6	6.3	420	7.7	16.1	61
Japan	29,251	4.4	4.5	6.4	198	9.7	10.7	65
Luxembourg	69,961	4.6	..	6.2	266	22	23.3	68
Netherlands	31,789	6.2	9.2	6.1	315	8.7	34.2	76
New Zealand	23,413	3.6	12.5	6.3	237	8.8	32.2	81
Norway	38,454	4.6	6.1	8.6	313	9.9	37.9	87
Portugal	19,629	7.5	15	6.7	342	5.6	21.3	79
Spain	25,047	9.1	10.3	5.5	330	7.3	30.5	65
Sweden	29,541	5.6	6.2	8	328	5.9	45.3	87
Switzerland	33,040	4.1	9	6.7	361	5.6	24.8	79
United Kingdom	30,821	4.8	13.8	6.9	230	9.4	18.5	79
United States	39,676	5.1	15.9	6.8	256	19.8	15	81

Source: hdr.undp.org/statistics/data
Unemp. = % Unemployed, Inequal. = Measure of inequality, Women parl. = % of seats in parliament held by women, Fem. econ. = Female economic activity (% of male rate).

a percentage of the male rate. These data are the "OECD data" file at the text Web site.

(a) Construct a stem-and-leaf plot of the GDP values, by rounding and reporting the values in thousands of dollars (e.g., replacing $19,629 by 20).

(b) Construct a histogram corresponding to the stem-and-leaf plot in (a).

(c) Identify the outlier in each plot.

3.7. Recently, the statewide number of abortions per 1000 women 15 to 41 years of age, for states in the Pacific region of the United States, were: Washington, 26; Oregon, 17; California, 236; Alaska, 2; and Hawaii, 6 (*Statistical Abstract of the United States, 2006*).

(a) Find the mean.

(b) Find the median. Why is it so different from the mean?

3.8. Global warming seems largely a result of human activity that produces carbon dioxide emissions and other greenhouse gases. The *Human Development Report 2005*, published by the United Nations Development Programme, reported per capita emissions in 2002 for the eight largest countries in population size, in metric tons (1000 kilograms) per person: Bangladesh 0.3, Brazil 1.8, China 2.3, India 1.2, Indonesia 1.4, Pakistan 0.7, Russia 9.9, United States 20.1.

(a) For these eight values, find the mean and the median.

(b) Does any observation appear to be an outlier? Discuss its impact on how the mean compares to the median.

3.9. A Roper organization survey asked, "How far have environmental protection laws and regulations gone?" For the possible responses not far enough, about right, and too far, the percentages of responses were 51%, 33%, and 16%.

(a) Which response is the mode?

(b) Can you compute a mean or a median for these data? If so, do so; if not, explain why not.

3.10. A researcher in an alcoholism treatment center, to study the length of stay in the center for first-time patients, randomly selects ten records of individuals institutionalized within the previous two years. The lengths of stay, in days, were 11, 6, 20, 9, 13, 4, 39, 13, 44, and 7.

(a) Construct a stem-and-leaf plot.

(b) Find the mean and the standard deviation, and interpret.

(c) For a similar study 25 years ago, lengths of stay for ten sampled individuals were 32, 18,

55, 17, 24, 31, 20, 40, 24, 15. Compare results to those in the new study using (i) a back-to-back stem-and-leaf plot, (ii) the mean, (iii) the standard deviation. Interpret any differences you find.

(d) Actually, the new study also selected one other record. That patient is still institutionalized after 40 days. Thus, that patient's length of stay is at least 40 days, but the actual value is unknown. Can you calculate the mean or median for the complete sample of size 11 including this partial observation? Explain. (This observation is said to be *censored*, meaning that the observed value is "cut short" of its true, unknown value.)

3.11. Access the GSS at sda.berkeley.edu/GSS. Entering TVHOURS for the variable and year(2006) in the selection filter, you obtain data on hours per day of TV watching in the U.S. in 2006.

(a) Construct the relative frequency distribution for the values 0, 1, 2, 3, 4, 5, 6, 7 or more.

(b) How would you describe the shape of the distribution?

(c) Explain why the median is 2.

(d) The mean is larger than 2. Why do you think this is?

3.12. Table 3.12 shows 2003 female economic activity (number of women in labor force per 100 men in labor force), for countries in Western Europe. Construct a back-to-back stem-and-leaf plot of these values contrasted with those from South America in Table 3.4. What is your interpretation?

3.13. According to Statistics Canada, in 2000 household income in Canada had median $46,752 and mean $71,600. What would you predict about the shape of the distribution? Why?

3.14. Table 3.13 summarizes responses of 2333 subjects in the 2006 General Social Survey to the question, "About how often did you have sex during the last 12 months?"

(a) Report the median and the mode. Interpret.

(b) Treat this scale in a quantitative manner by assigning the scores 0, 0.1, 1.0, 2.5, 4.3, 10.8, and 17 to the categories, for approximate monthly frequency. Find the sample mean, and interpret.

TABLE 3.13

How Often Had Sex	Frequency
Not at all	595
Once or twice	205
About once a month	265
2 or 3 times a month	361
About once a week	343
2 or 3 times a week	430
More than 3 times a week	134

3.15. The 2004 GSS asked respondents "How often do you read the newspaper?" The possible responses were (every day, a few times a week, once a week, less than once a week, never), and the counts in those categories were (358, 222, 134, 121, 71).

(a) Identify the mode and the median response.

(b) Let y = number of times you read the newspaper in a week, measured as described above. For the scores (7, 3, 1, 0.5, 0) for the categories, find $\bar{y}$. How does it compare to the mean of 4.4 for the 1994 GSS?

3.16. According to the U.S. Bureau of the Census, *2005 American Community Survey*, the median earnings in the past 12 months was $32,168 for females and $41,965 for males, whereas the mean was $39,890 for females and $56,724 for males.

(a) Does this suggest that the distribution of income for each gender is symmetric, or skewed to the right, or skewed to the left? Explain.

(b) The results refer to 73.8 million females and 83.4 million males. Find the overall mean income.

TABLE 3.12

Country	Female Econ. Activity	Country	Female Econ. Activity	Country	Female Econ. Activity
Austria	66	Germany	71	Norway	86
Belgium	67	Greece	60	Portugal	72
Cyprus	63	Ireland	54	Spain	58
Denmark	85	Italy	60	Sweden	90
Finland	87	Luxembourg	58	U.K.	76
France	78	Netherlands	68		

Source: *Human Development Report, 2005*, United Nations Development Programme.

3.17. In 2003 in the United States, the median family income was $55,800 for white families, $34,400 for black families, and $34,300 for Hispanic families (*Statistical Abstract of the United States, 2006*).
- **(a)** Identify the response variable and the explanatory variable for this analysis.
- **(b)** Is enough information given to find the median when all the data are combined from the three groups? Why or why not?
- **(c)** If the reported values were means, what else would you need to know to find the overall mean?

3.18. The GSS has asked, "During the past 12 months, how many people have you known personally that were victims of homicide." Table 3.14 shows a printout from analyzing responses.
- **(a)** Is the distribution bell shaped, skewed to the right, or skewed to the left?
- **(b)** Does the Empirical Rule apply to this distribution. Why or why not?
- **(c)** Report the median. If 500 observations shift from 0 to 6, how does the median change? What property does this illustrate for the median?

3.19. As of October 2006, an article in wikipedia.org on "Minimum wage" reported (in U.S. dollars) the minimum wage per hour for five nations: $10.00 in Australia, $10.25 in New Zealand, $10.46 in France, $10.01 in the U.K., $5.15 in the U.S. Find the median, mean, range, and standard deviation **(a)** excluding the U.S., **(b)** for all five observations. Use the data to explain the effect of outliers on these measures.

3.20. *National Geographic Traveler* magazine recently presented data on the annual number of vacation days averaged by residents of eight different countries. They reported 42 days for Italy, 37 for France, 35 for Germany, 34 for Brazil, 28 for Britain, 26 for Canada, 25 for Japan, and 13 for the United States.
- **(a)** Find the mean and standard deviation. Interpret.
- **(b)** Report the five-number summary. (*Hint:* You can find the lower quartile by finding the median of the four values below the median.)

3.21. The Human Development Index (HDI) is an index the United Nations uses to give a summary rating for each nation based on life expectancy at birth, educational attainment, and income. In 2006, the ten nations (in order) with the highest HDI rating, followed in parentheses by the percentage of seats in their parliament held by women (which is a measure of gender empowerment) were Norway 38, Iceland 33, Australia 28, Ireland 14, Sweden 45, Canada 24, Japan 11, United States 15, Switzerland 25, Netherlands 34. Find the mean and standard deviation, and interpret.

3.22. The *Human Development Report 2006*, published by the United Nations (UN), showed life expectancies by country. For Western Europe, the values reported were

> Denmark 77, Portugal 77, Netherlands 78, Finland 78, Greece 78, Ireland 78, UK 78, Belgium 79, France 79, Germany 79, Norway 79, Italy 80, Spain 80, Sweden 80, Switzerland 80.

For Africa, the values reported (many of which were substantially lower than five years earlier because of the prevalence of AIDS) were

> Botswana 37, Zambia 37, Zimbabwe 37, Malawi 40, Angola 41, Nigeria 43, Rwanda 44, Uganda 47, Kenya 47, Mali 48, South Africa 49, Congo 52, Madagascar 55, Senegal 56, Sudan 56, Ghana 57.

- **(a)** Which group of life expectancies do you think has the larger standard deviation? Why?
- **(b)** Find the standard deviation for each group. Compare them to illustrate that s is larger for the group that shows more spread.

TABLE 3.14

VICTIMS	Frequency	Percent
0	1244	90.8
1	81	5.9
2	27	2.0
3	11	0.8
4	4	0.3
5	2	0.1
6	1	0.1

N	Mean	Std Dev	Max	Q3	Med	Q1	Min
1370	0.146	0.546	6	0	0	0	0

3.23. A report indicates that teacher's annual salaries in Ontario have a mean of $50,000 and standard deviation of $10,000 (Canadian dollars). Suppose the distribution has approximately a bell shape.
(a) Give an interval of values that contains about (i) 68%, (ii) 95%, (iii) all or nearly all salaries.
(b) Would a salary of $100,000 be unusual? Why?

3.24. Excluding the U.S., the national mean number of holiday and vacation days in a year for OECD nations (see Exercise 3.6) is approximately bell shaped with a mean of 35 days and standard deviation of 3 days.[1]
(a) Use the Empirical Rule to describe the variability.
(b) The observation for the U.S. is 19. If this is included with the other observations, will the (i) mean increase, or decrease, (ii) standard deviation increase, or decrease?
(c) Using the mean and standard deviation for the other countries, how many standard deviations is the U.S. observation from the mean?

3.25. For GSS data on "the number of people you know who have committed suicide," 88.8% of the responses were 0, 8.8% were 1, and the other responses took higher values. The mean equals 0.145, and the standard deviation equals 0.457.
(a) What percentage of observations fall within one standard deviation of the mean?
(b) Is the Empirical Rule appropriate for the distribution of this variable? Why or why not?

3.26. The first exam in your Statistics course is graded on a scale of 0 to 100, and the mean is 76. Which value is most plausible for the standard deviation: −20, 0, 10, or 50? Why?

3.27. Grade point averages of graduating seniors at the University of Rochester must fall between 2.0 and 4.0. Consider the possible standard deviation values: −10.0, 0.0, 0.4, 1.5, 6.0.
(a) Which is the most realistic value? Why?
(b) Which value is *impossible*? Why?

3.28. According to the U.S. Census Bureau, the U.S. nationwide median selling price of homes sold in 2005 was $184,100. Which of the following is the most plausible value for the standard deviation: **(a)** −15,000, **(b)** 1,000, **(c)** 10,000, **(d)** 60,000, **(e)** 1,000,000? Why?

3.29. For all homes in Gainesville, Florida, the residential electrical consumption[2] for the year 2006 had a mean of 10,449 and a standard deviation of 7489 kilowatt-hours (kWh). The maximum usage was 336,240 kWh.
(a) What shape do you expect this distribution to have? Why?
(b) Do you expect this distribution to have any outliers? Explain.

3.30. Residential water consumption (in thousands of gallons) in Gainesville, Florida in 2006 had a mean of 78 and a standard deviation of 119. What shape do you expect this distribution to have? Why?

3.31. According to *Statistical Abstract of the United States 2006*, mean salary (in dollars) of secondary school teachers in 2004 in the United States varied among states with a five-number summary of

```
100% Max      61,800 (Illinois)
 75% Q3       48,850
 50% Med      42,700
 25% Q1       39,250
  0% Min      33,100 (South Dakota)
```

(a) Find and interpret the range.
(b) Find and interpret the interquartile range.

3.32. Refer to the previous exercise.
(a) Sketch a box plot.
(b) Based on (a), predict the direction of skew for this distribution. Explain.
(c) If the distribution, although skewed, is approximately bell shaped, which value is most plausible for the standard deviation: (i) 100, (ii) 1000, (iii) 7000, (iv) 25,000? Explain.

3.33. Table 3.15 shows part of a computer printout for analyzing the murder rates (per 100,000) in the "2005 statewide crime" data file at the text Web site. The first column refers to the entire data set, and the second column deletes the observation for D.C. For each statistic reported, evaluate the effect of including the outlying observation for D.C.

3.34. During a recent semester at the University of Florida, computer usage[3] of students having accounts on a mainframe computer was summarized by a mean of 1921 and a standard deviation of 11,495 kilobytes of drive usage.
(a) Does the Empirical Rule apply to this distribution? Why?
(b) The five-number summary was minimum = 4, $Q1$ = 256, median = 530, $Q3$ = 1105, and maximum = 320,000. What does this suggest about the shape of the distribution? Why?
(c) Use the 1.5(IQR) criterion to determine if any outliers are present.

[1] *Source*: Table 8.9 in www.stateofworkingamerica.org, from The Economic Policy Institute.
[2] Data supplied by Todd Kamhoot, Gainesville Regional Utilities.
[3] Data supplied by Dr. Michael Conlon, University of Florida.