

# 2

## EXPLORING DISTRIBUTIONS

### Overview

---

#### Goals

The overall goal of Chapter 2 is to provide a systematic way to uncover information through displaying and summarizing distributions of univariate data. Students will get a sound knowledge of the basic concepts and tools for exploring distributions and an understanding of how the tools compare. If students merely memorize the set of graphical and numerical techniques, they will have missed the point. The type of information uncovered through exploratory techniques should prompt students to ask questions that can be answered by the inferential techniques in later chapters.

In this chapter, students will learn

- to make and interpret plots for displaying univariate data: dot plot, histogram, stemplot, and boxplot
- to compute and interpret measures of center: mean and median
- to compute and interpret measures of spread: interquartile range and standard deviation
- to examine the effects of a linear transformation of the data on measures of center and spread
- to understand the use of normal density curves as models for data distributions
- to calculate probabilities connected with the normal distribution using standard units ( $z$ -scores) and a table of the standard normal distribution

## Content Overview

Understanding what data represents is most easily seen by clearly distinguishing between *variables* and *cases*. A variable is that feature being measured, and a case is that object or person on which the measurement is made. In the spreadsheet data in Display 2.24, the variables are the column headings, (gestation period, average longevity, etc.) and the cases are the row headings (baboon, grizzly bear, etc.). For each case (type of mammal), measurements are made of gestation period, average longevity, and so forth.

### Types of Variables

*Categorical variables* specify the group (category) to which a case belongs. Whether a mammal is considered wild or not and whether a mammal is considered to be a predator or not are the two categorical variables in Display 2.24. Coding wild as “1” and domestic as “0” is convenient but arbitrary. We could just as well have used “W” and “D” rather than “1” and “0.” The 0–1 coding is convenient because the sum of the Wild column gives the number of mammals that are wild.

A *quantitative variable* assigns a numerical value or measurement to each case. Gestation period (in days), average longevity (in years), maximum longevity (in years), and speed (in mph) are the quantitative variables in Display 2.24.

### Displaying Distributions

The set of values a variable takes on across the cases, along with the frequency with which each value occurs, produces a *distribution* of the data. Display 2.33 shows the distribution of the categorical variable describing whether a mammal is wild or not. Display 2.29 shows the distribution of the quantitative variable of mammal speeds. In fact, almost every plot in Chapter 2 displays a distribution of a single variable. Distributions are the key to statistical investigations. (Later, probability distributions and sampling distributions will be added to the array of distributions essential for statistics.) **One of the first steps in understanding statistics is learning to think in terms of distributions and where the value that is of interest to you fits into the appropriate distribution.** A picture may be worth a thousand words, but a statistical graphic is worth even more. Often it is the only way to see essential properties of a distribution. Thus, modern statistics is highly graphical. So stress the importance of graphical summaries and displays in the process of understanding distributions.

These are the standard plots used to display a single variable:

#### Categorical Variables

Bar graph

#### Quantitative Variables

Dot plot

Stemplot (stem-and-leaf plot)

Histogram

Cumulative frequency plot

Boxplot (box-and-whiskers plot)

For categorical variables:

**Bar graph:** A bar graph shows the frequencies (or relative frequencies) for the various categories of a categorical variable. Display 2.33 is a bar graph of the distribution of wild versus domestic mammals in the data set of Display 2.24. Note that the bars do not have to be arranged in any particular order because the horizontal axis is not a numerical scale. (See also Display 2.35.)

For quantitative variables:

**Dot plot:** A dot plot has a small dot above a real number line to mark the value for each case. Section 2.1 introduces students to distributions and their properties by using easily understood dot plots, which were introduced in Chapter 1. Dot plots quite nicely show the shape of a distribution, and you can use them for either large or small data sets.

**Stemplot or stem-and-leaf plot:** Display 2.29 gives a stemplot of the mammal speeds. Stemplots have the advantage of preserving at least two digits of the actual numerical values of the data. They are most useful for data sets of small to moderate size and can be made quickly by hand.

**Histogram:** A histogram places a bar above intervals of values of a variable, with the height of a bar indicating the frequency (or relative frequency) of the cases for that interval. Display 2.26 shows a histogram of the mammal speed data. Histograms do not preserve the exact numerical values in the data set. They are most useful for data sets of moderate to large size and are best made by a computer.

**Cumulative frequency plot:** A cumulative frequency plot (or cumulative relative frequency plot) shows the number (or proportion) of the values in a data set that are a given size or smaller. See Display 2.55.

**Boxplot:** A five-number summary consisting of the extremes, the quartiles, and the median can be plotted on a real number line, with the quartiles connected by a box, to form a boxplot or *box-and-whiskers plot* for the data. Boxplots do not give detailed information on shape, as found in other plots described above, but do provide a quick summary of location, spread, and possible skewness. See Display 2.50.

These plots all give some indication of the shape, center, and spread of the data. You should choose the one (or two) that best meets the needs of the situation. In most cases, you should try more than one to see which works best.

## Describing Distributions

Here are the key strategies for describing distributions.

For categorical variables: Focus on the frequency or relative frequency. The data for categorical variables, as pictured in a bar graph, are summarized in a table of frequencies. Categorical variables will be covered in Chapters 6, 8, and 10.

For quantitative variables: Focus on shape, center, and spread.

**Shape** is described in terms of general shape, *symmetry* versus *skewness*, and in terms of the possible presence of *outliers*, *clusters*, and *gaps*. These terms, with the possible exception of *outlier*, are only loosely defined, and a great deal of judgment can be necessary when describing the shape of a distribution.

**Center** is a location for the data distribution on the real number line. The center is at the line of symmetry for a symmetric distribution, but more options present themselves for skewed distributions. The two common measures of center are the *median* and the *mean* (or average), both of which lie at the line of symmetry for symmetric distributions. For skewed distributions, the median will lie close to the bulk of the data points, whereas the mean will lie closer to the tail. The *mode*, or most common value, is the location of the highest bar on a histogram. It will play almost no part in this book, except that distributions with two peaks will be called *bimodal*.

**Spread** is a measure of the variability in a data set. If the median is an appropriate measure of center, then the distance between the quartiles (*interquartile range*) is typically an appropriate measure of spread. If the quartiles are close to the median, there is little variability in the data; if at least one of the quartiles is far from the median, there is considerable variability in the data. If the mean is an appropriate measure of center, then the standard deviation is typically an appropriate measure of spread. The standard deviation measures the “typical” distance between the values and their mean. It is a good measure of the typical deviation from the mean when the data distribution has a single peak and is reasonably symmetric.

### **A Systematic Approach to Exploring Univariate Data**

Any data exploration should follow the steps of

plot → shape → center → spread

That is, choose an appropriate plot, describe the shape, and find a measure of center appropriate to the shape and a measure of spread that agrees with the measure of center.

Ultimately, which measure of center and spread to use depends on the purpose for computing a summary statistic. If the purpose is simply to describe a distribution, it is most informative to use the mean and standard deviation for distributions that are approximately normal in shape and the median and interquartile range for skewed distributions.

### **In the Final Analysis**

Being able to choose appropriate graphical and numerical summaries of data and being able to write or verbalize a coherent summary of what the data show is more important than mastering details of computation that a calculator or computer can handle.

## **Instructional Methods**

Because one goal of this chapter is to have students learn to display and summarize sets of data, even large sets of data, it is important that they have practice doing just that. Now is the time for students to begin learning how to use their graphing calculators and statistical software. Statistical software enables students to construct plots quickly, accurately, and flexibly.

Statistics must be taught using data from the real world, so it is important that you review the discussion, practice, and exercise items for their suitability for your class. Statistics is an incredibly powerful tool in the social sciences and in the life sciences, and this text draws from many areas of these disciplines. Consequently, some contexts may be too sensitive for some of your students. Only you can make that decision. We recommend that you routinely review the social contexts of the material before presenting it.

Move through this chapter as quickly as possible. The amount of time you’ll need will depend on how much statistics your students have previously learned in their high school mathematics curriculum.

Students may not become thoroughly comfortable with several of the concepts in this chapter, such as standard deviation and transforming data. These ideas will be revisited in later chapters, and by the end of the course students will have mastered them.

## Time Required

Traditional Schedule			Block	4 x 4 Block
<b>Section 2.1</b>				
1–2 days	Day 1	Overview, Activity 2.1, uniform, normal	1.5 days	1 long, 1 short
	Day 2	Skewed, bimodal, summary, exercises		
<b>Section 2.2</b>				
1–2 days	Day 1	Dot plot, histogram, stemplot	1.5 days	1 long, 1 short
	Day 2	Activity 2.2, bar graph, summary, exercises		
<b>Section 2.3</b>				
5–6 days	Day 1	Mean, median	5 days	2 long, 2 short
	Day 2	Quartiles, five-number summaries, boxplots, modified boxplots		
	Day 3	Percentiles, cumulative frequency plots		
	Day 4	Activity 2.3, standard deviation		
	Day 5	Properties of summary statistics, recentering and rescaling		
	Day 6	Influence of outliers, summaries from a frequency table, summary, exercises		
<b>Section 2.4</b>				
2–3 days	Day 1	Unknown value and percentage problems, standard normal curve	2 days	1 long, 2 short
	Day 2	Determining z-scores, solving unknown value and percentage problems		
	Day 3	Central intervals, summary, exercises		
<b>Review</b>				
1 day			1 day	1 day

## Materials

**Section 2.1:** For Activity 2.1, a tennis ball (a dead one is fine) and a centimeter ruler for every two students

**Section 2.2:** For Activity 2.2, a yardstick or meterstick

**Section 2.3:** For Activity 2.3, a ruler for each group of four students

**Section 2.4:** None

## Suggested Assignments

Classwork			
Section	Essential	Recommended	Optional
2.1	Activity 2.1 D1, D2, D4 P1, P2, P4–P6	D3a P3, P7	D3b (must do Activity 2.1 before D3b), D5–D7
2.2	D8, D12–D15, D17 P9–P11, P13–P15	D9, D10, D16 P8, P12, P16	Activity 2.21 D11
2.3	D18, D20, D22–D30, D33–D36 P17, P19–P21, P23–P25, P27, P29, P30, P32, P34, P35	Activity 2.3 D19, D31, D37 P18, P22, P28, P31, P33, P36	D21, D32 P26
2.4	D39, D40, D42, D43, D45, D46 P38–P42, P44, P46–P48	D38, D41, D44 P37, P43, P45	

Homework			
Section	Essential	Recommended	Optional
2.1	E1–E5	E6, E7, E10, E11	E8, E9
2.2	E12, E16, E20	E13, E14, E19, E22	E15, E17, E18, E21
2.3	E24, E26–E29, E33, E35, E40	E23, E25, E30, E37, E38, E41–E43	E31, E32, E34, E36, E39, E44
2.4	E46–E48, E52, E54, E56	E45, E49–E51, E53, E55	
Review	E57–E59, E62, E63, E65, E67, E69, E71	E60, E66, E74	E61, E64, E68, E70, E72, E73, E75, E76

## 2.1 The Shapes of Things: Visualizing Distributions

---

### Objectives

- to learn the basic shapes of distributions of data—uniform, normal, skewed
- to describe the characteristics of the shape of a distribution, including symmetry, skewness, modes, outliers, gaps, and clusters
- to describe a uniform distribution using the range and the frequency
- to estimate graphically the mean and standard deviation of a normal distribution and use them to describe the distribution
- to estimate graphically the median and quartiles and use them to describe a skewed distribution

Mastery of the mean, median, quartiles, and standard deviation is not expected until later sections of this chapter.

### Important Terms and Concepts

- *basic shape of a distribution*: rectangular or uniform, normal, skewed right, skewed left
- *characteristics of the shape of a distribution*: symmetric, skewed, bimodal, outliers, gaps, clusters
- *measures of center*: mean, median
- *measures of spread*: standard deviation, quartiles

### Lesson Planning

#### Class Time

One to two days. We suggest that you move quickly through this section.

#### Materials

For Activity 2.1, a tennis ball (a dead one is fine) and a centimeter ruler for every two students

## Suggested Assignments

Classwork		
Essential	Recommended	Optional
Activity 2.1* D1, D2, D4 P1, P2, P4–P6	D3a P3, P7	D3b (must do Activity 2.1 first), D5–D7
*This activity is highly recommended because it demonstrates to students the variability inherent in even careful measurement of the same thing.		
Homework		
Essential	Recommended	Optional
E1–E5	E6, E7, E10, E11	E8, E9

### Lesson Notes: The Shapes of Things

The dot plots in this section were made by statistical software. Because of the limitations of the resolution of the computer screen, the dot plot in Display 2.3, for example, has places for 10 vertical lines of dots on which to plot the values from 0 through 0.20. Students don't need to worry about the exact rounding rule used by the software.

#### Activity 2.1: Measuring Diameters

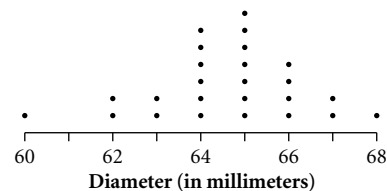
Activity 2.1 introduces students to the concept of measurement error. The term *error* is not used in the sense of “mistake” but rather in the sense of “variation.”

Throughout the course, students will gather data and combine their individual responses into a large data set. Make sure that students understand that variation generally occurs in the process of gathering data and may not represent an error on the part of anyone associated with the process. Variation occurs from one measurement of an object to another measurement of the same object, and from one object to another object. Variation also occurs in chance processes. For example, if you flip a fair coin, you expect that you will get 50% heads and 50% tails. However, you may get 6 heads out of 10 flips or 29 heads out of 50 flips just by chance.

1. Methods will vary. Some students will hold the tennis ball between two books and measure the distance between the books. Others might wrap a string around the ball, measure the length of the string, and then divide by  $\pi$ .

2. Answers will vary. Most measurements should be between about 60 and 70 mm.

3. This plot and print-out summarize the data for 25 student measurements (in millimeters) of the diameters of tennis balls. Each pair of students had their own ball, so there may be a little ball-to-ball variation. Most of the variation, however, is due to measurement error; it is difficult to measure the diameter of a sphere.



	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
Diameter	25	64.640	65.000	64.652	1.655	.331
	MIN	MAX	Q1	Q3		
Diameter	61.000	68.000	64.000	66.000		

4. Like the one in step 3, distributions should be somewhat symmetrical and mound-shaped, which is typical of distributions of measurement errors.

5. Answers will vary. A typical result would be “about 65 mm, give or take about 2 mm or so.”

6. Sources of variability include differences from tennis ball to tennis ball, differences in method of measurement, and error in reading or placing the centimeter scale. The variability could be reduced by using just one tennis ball, by all groups using the

same method of measuring, and by all groups being trained in the method before starting. The variability cannot be eliminated entirely. As an example of why, suppose the diameter actually was 65.5 mm. When measuring to the nearest millimeter, some groups would report 65 and others 66 depending on whether they judged the measurement to be slightly under or slightly over 65.5 mm.

## Lesson Notes: Uniform or Rectangular Distribution

The area over the closed interval  $0.2 \leq x \leq 0.4$  or  $[0.2, 0.4]$  is the same as over the open interval  $0.2 < x < 0.4$  or  $(0.2, 0.4)$  because the area over a single point is 0.

### Discussion

As in Chapter 1, discussion questions were written to guide you in leading a class discussion of the previous material. You don't need to cover every one or cover them in order if your students lead the discussion in a different direction. Many questions will require you to provide significant help. The questions weren't designed for students to write out answers, except perhaps some notes to use in studying for exams.

**D1.** Answers will vary for this question. For part a, for example, each answer should involve the number of occurrences or frequency of some phenomenon that is attached to the days of the week. The important thing here is to discuss each suggested answer so that students learn what constitutes a distribution and what doesn't. For part a, students might suggest the temperature on each day of the week. That wouldn't be a distribution because the value on the  $y$ -axis would not be a frequency. However, a good example might be the number of days in the past year that the temperature rose above 70 degrees for each day of the week. That generates a distribution that is most likely uniform. Most suggestions will generate some discussion—and that's the idea. For example, a student might suggest that the number of, say, visits to McDonald's would have a uniform distribution over the days of the week. With class discussion, he or she should soon realize that there are more visits on Saturday than on, say, Wednesday, so this would not be a good example.

**a.** A good example would involve something that is equally likely to occur on each day of the week, such as the birthdays of classmates or the number of shooting stars seen in the night sky.

**b.** Possible examples are the last digits of phone numbers of students in the class or the last digit of the height in millimeters of members of the class.

**D2. a.** Possible examples include the number of automobile accidents or the number of visits to the doctor because both occur more frequently in winter in most places. Shooting stars are more likely in some months than others, a notable meteor shower occurring each August. Because February has fewer days, almost anything should occur less frequently in February.

**b.** A possible example is the number of bills paid each day of the month in a given family, which tend to be clustered around given days of the month, such as payday.

**c.** Possible examples include the frequency of students in the class with that digit as the last digit of their age or the frequency of days in a year that the traffic deaths in a small town reached that number.

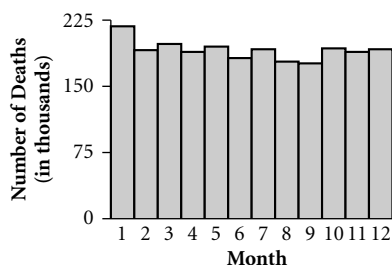
**d.** Possible examples include the number of trips to the beach a person makes in a lifetime on that day or the number of people who attend the local movie theater on that day.

### Practice

As in Chapter 1, practice questions are designed for students to do individually to make sure each student understands the preceding material. Usually, these will be done as "seat work"; other times, they will be part of the homework.

**P1.** An example plot follows. The number of deaths per month is fairly constant across the months, with about 190,000 per month. The exception is January, which shows a higher number of deaths than the other months.

The up-and-down nature of the plot appears to be a result of the fact that some months have more days than others. However, the up-and-down pattern should have been broken from July to August, which have the same number of days.



- P2. a. With a perfect uniform distribution on  $[0, 2]$ , the value 1.0 would divide the values in half.  
 b. 0.5, 1.0, and 1.5  
 c. 0.5 and 1.5  
 d. 0.15  
 e. 0.05 and 1.95

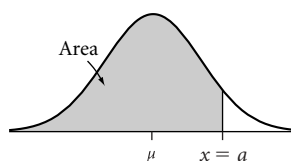
## Lesson Notes: Normal Distribution

The term *normal* should not be applied casually to any distribution that is mound-shaped and symmetric. Instead, use terms like *round-shaped*, *approximately normal*, *bell-shaped*, and so forth to describe the shape of a data distribution. The normal distribution has a precise definition, and the equation of a normal curve is

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5\left(\frac{x-\mu}{\sigma}\right)^2}$$

This equation is completely determined by the two parameters of the distribution—the mean,  $\mu$ , and standard deviation,  $\sigma$ . The integral below gives the area under this curve below a specified value  $a$  on the  $x$ -axis.

$$\text{area} = \int_{-\infty}^a \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5\left(\frac{x-\mu}{\sigma}\right)^2} dx$$



This function does not have a closed-form anti-derivative, so the definite integral can be evaluated only by numerical methods.

The section on the normal distribution on pages 27 to 29 contains an introduction to the standard deviation. The approach here is entirely graphical. Students learn that the distance from the center of a

normal curve to either inflection point is one standard deviation and that the region within one standard deviation of the mean contains about two-thirds of the values. This approach fits with their interpretation of other measures of spread:

**Range:** The interval between the minimum and maximum contains 100% of the data.

**IQR:** The interval between the first and third quartiles contains 50% of the data.

**SD:** The interval between one *SD* below and one *SD* above the mean contains two-thirds of the data if the distribution is normal.

No formula is given at this time for the standard deviation. We have found that if the formula is given too early, students concentrate on questions like “Why do we divide by  $n - 1$  rather than by  $n$ ?” and “Why are we squaring things?” and “Where did the square root come from?” Try to get students to think graphically about the standard deviation at this stage so they will be comfortable with it as a measure of spread before encountering the formula in Section 2.3.

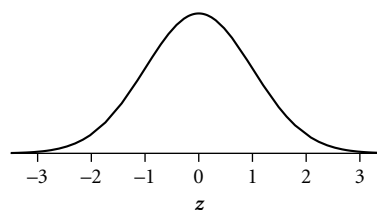
The region within one standard deviation of the mean actually encloses closer to 68% of the values than two-thirds, which you may want to tell your students at this point or wait until this comes up in Section 2.4.

### Discussion

- D3. a. The weight of a typical penny is about 3.11 grams, give or take 0.04 grams or so.  
 b. Answers will vary depending on measurements from your class. The mean should be about 65 mm and the standard deviation about 1.7 mm.

### Practice

P3.



- P4. Student estimates will differ somewhat from the actual means and standard deviations given here.  
 a. A typical SAT verbal score is roughly 500, give or take about 100 or so.

- b. A typical ACT score is about 20, give or take 5 or so.
- c. A typical college-aged woman is about 65 inches tall, give or take 2.5 inches or so.
- d. A typical professional baseball player in the 1910s had a single-season batting average of about .260 or .270, give or take about .040 or so.

## Lesson Notes: Skewed Distribution

In this section, students will be estimating quartiles graphically, not computing them. Computation will begin in Section 2.3. For now, have students think graphically about quartiles so they understand the concept.

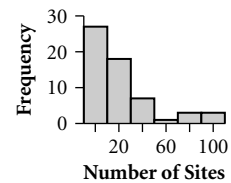
### Discussion

- D4.**
- a. This distribution is strongly skewed right. Most islands are quite small; Cuba and Hispaniola are comparatively very large. There is a wall at 0 because no island can be smaller than that, but many are close.
  - b. This distribution should be skewed right because some countries, such as the United States, have a per capita income that is much higher than most other countries. There is a wall near 0 because average per capita incomes can't go below that, but for many countries the per capita income is very small.
  - c. This distribution should not be skewed at all. Lengths like this typically form a distribution that is approximately normal.
  - d. This distribution probably will be skewed left. There is a wall at 1 hour—no student can take longer than that, and most students will work on an exam for the entire hour or close to it. A few students, however, will leave early.
  - e. This distribution will be skewed right. Some emperors reigned a long time but most for a moderate number of years. There is a wall at 0 years.
- D5.** Variables that tend to have a few large values and many relatively small ones include sizes of corporations (in either dollars or people) and land areas of the states. A variable that is actually a maximum also tends to be skewed right, such as maximum speeds of different models of cars or the most expensive shirt in each of the stores in a large mall.

- D6.** Variables that tend to have many large values and a few relatively small ones include scores on an easy exam, ages of residents of a retirement home, and anything that is actually a minimum, like the lowest priced shirt in each of the stores in a large mall.
- D7.** Toward the right, because it is very common to have a wall at 0 because most quantities must be positive.

### Practice

- P5.** a. IV      b. II      c. V  
d. III      e. I
- P6.** Students should make a plot that is skewed right. Here is a histogram of the actual distribution:



- P7.** There are 61 GPAs, so about 15 should fall into each quarter. The lower quartile is about 2.9, the median about 3.35, and the upper quartile about 3.7. The middle 50% of the students had GPAs between 2.9 and 3.7, with half above 3.35 and half below. (If the class were made up of older students, the distribution of GPAs might shift away from 4.0 and become less skewed. Students with more classes to their credit will have a harder time maintaining a high GPA.)

## Lesson Notes: Bimodal Distribution and Other Features: Outliers, Gaps, and Clusters

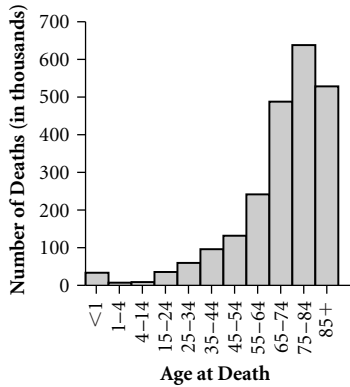
In Section 2.3, students will get a rule to use to identify a value as an outlier. At this point, the purpose of introducing the term is to provide more tools for students to use in describing distributions. Although the mode is not an important summary statistic, the idea of *bimodal* as a description of a distribution should be in the AP Statistics student's vocabulary.

There are no discussion problems or practice problems provided for these two short sections. If you have time, you could use E10 and E11 for that purpose.

## Exercises

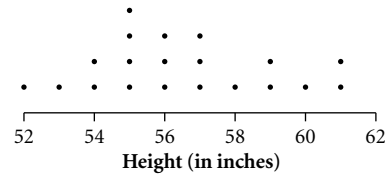
Exercises were designed for students to do as homework. Most can be done either individually or in groups.

- E1. a.** This distribution is strongly skewed left. The actual distribution follows. Students will often have the height of the bar for 85+ taller than that for 75–84, confusing actual number of deaths with probability of death. There are fewer people in the 85+ category than in the 75–84 category, so fewer of them die.



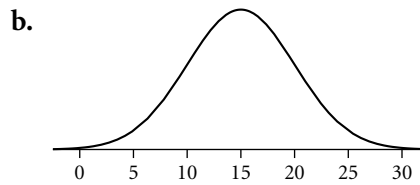
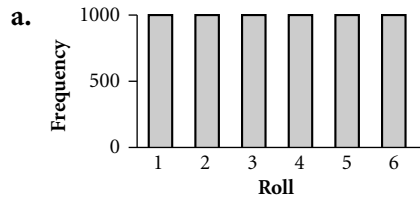
Source: *Statistical Abstract of the United States, 1997, Table 130.*

- b.** This distribution will be strongly skewed right. Most people get their driver's licenses at the earliest possible age or quite close to it.
- c.** The distribution of SAT scores for a large number of students should be approximately normal.
- d.** Selling prices of new cars should show a few very expensive models (like Corvettes) and a large number of relatively inexpensive (but not cheap!) ones (around \$15,000 to \$20,000). The distributions should be skewed toward the larger values.
- E2. a.** skewed right (toward larger values)
- b.** bimodal; developing countries tend to have higher birth rates than do developed countries.
- c.** Approximately normal is a good answer, but in fact the distribution is slightly skewed right. The dot plot shown next gives the heights in inches of the U.S. women's soccer team that won the World Cup in 1999.

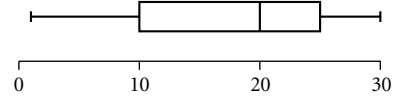


- d.** roughly uniform
- e.** skewed left (toward smaller values)

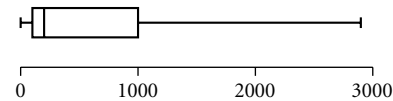
**E3.**



- c.** There are many possible answers. One such set of numbers is 1, 2, 3, 4, 9, 11, 13, 14, 17, 19, 21, 22, 23, 23, 24, 26, 27, 28, 29, 30. This boxplot represents one possible sketch.



- d.** There are many possible answers. One such set of numbers is 0, 10, 20, 30, 80, 120, 150, 160, 170, 190, 210, 400, 500, 600, 700, 1300, 1500, 1800, 2400, 2900. This boxplot represents one possible sketch.

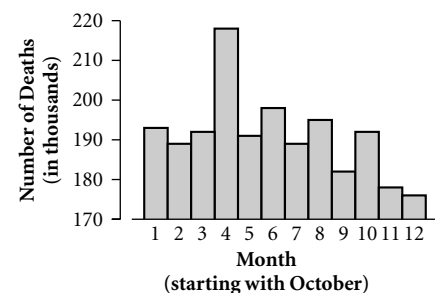
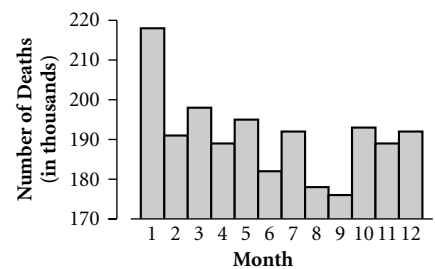


- E4.** The last digits of social security numbers are essentially random digits, so they should be fairly uniform over the range 0 to 9, as these are. Here we would say that the distribution is approximately uniform with about six students with each digit from 0 to 9.

- E5.** **a.** A case is one of the approximately 92 officers who attained the rank of colonel in the Royal Netherlands Air Force. There is only one variable: the age at which the officer became a colonel.
- b.** This distribution is skewed left with no outliers, gaps, or clusters. The median is 52 years, and the quartiles are 50 and 53. So we would say that the middle half of the ages are between 50 and 53, with half above 52 and half below.
- c.** In questions like this, students aren't expected to "guess" the "right" answer. Instead, they are expected to generate many possibilities that could then be investigated. For example, some military services have an "up or out" rule. It may be the case in the Royal Netherlands Air Force that if you haven't been promoted to colonel by your 55th birthday, you must retire. The armed services have very little use for 55-year-old privates. Alternatively, there might be a mandatory retirement at age 55. A third possibility is age discrimination against older people in the service.
- E6.** **a.** A case is one of the 125 seasons. The only variable is the number of inches of rain. Note: Rainfall is given by "seasons" rather than by years because Los Angeles has almost no rain from May through October. The rainy season is late fall through early spring. To give rainfall by year would be to break each of these natural seasons into two parts.
- b.** This distribution is skewed right with no obvious outliers. The three peaks suggest three groups of data points; perhaps weather conditions cause dry, normal, or rainy years. The median is 13, and the quartiles are 10 and 19. So we say that half the values are above 13 and half are below, with the middle half between 10 and 19.
- c.** The number of inches of rain cannot go below 0, so that is a natural wall. However, it appears that about 4 inches of rain is the effective minimum. There may be some characteristic of the weather that makes it almost impossible to go below that.
- E7.** The distribution is approximately normal, except that it has too many outliers and is a bit too "peaked" to be traced by a normal

curve. Because it is roughly symmetric, we can say that the distribution is centered at about 98.

- E8.** The up-and-down pattern looks fairly random, but we now see that births tend to be more frequent in the summer (July, August, and September). You may want to mention to students that they should always check plots to see where the scale begins. Although this plot shows more detail, if someone did not notice that the plot has been cut off at the bottom, they would think that the variation from month to month is larger than it actually is.
- E9.** The plot appears next (the first plot) with a scale beginning at 170. In this plot, January looks even more unusually high. As mentioned in P1, the up-and-down nature of the plot appears to be a result of the fact that some months have more days than others. However, the up-and-down pattern should have been broken from July to August, which have the same number of days. Now we can see the additional fact that the number of deaths seems to be generally declining over the months until October when it goes back up. Apparently, there are more deaths in the colder months. This becomes especially clear if we let October be month #1 and September be month #12, as was done in the second plot.



**E10. a.** The Nielsen data consist of 101 cases—the 101 television shows. Each case has two variables associated with it—the number of viewers who watched each show and the network on which it was shown. The number of viewers is a quantitative variable because it is a measured quantity or value. The network is a categorical variable because each show falls into exactly one of six categories (ABC, CBS, FOX, NBC, UPN, or WB).

**b.** As is often the case, the basic shape of the distribution of Nielsen ratings isn't clear-cut. The distribution is somewhat skewed right as the values rise rapidly from the left and taper off toward the right. There is a wall at 0—a show cannot have a negative number of viewers, no matter how awful it is.

The values in this distribution are all clumped together, with the exception of three high values for *Seinfeld*, *Seinfeld Clips*, and *ER*. These three highest values are widely separated from the bulk of the ratings. There are no gaps or clusters other than those created by the three highest values.

**c.** The median is 10.15. The spread is large, especially if you consider the three highest ratings. However, this large range is due to an outlier—*Seinfeld*. The middle half of the shows are fairly close together, grouped between ratings of about 6 and 13. (Actual values are 6.18 and 12.78.)

**d.** It had about seven times as many viewers as a typical show.

**e.** First, note that the scales on the two dot plots are different. The distribution for the regular week fits none of the basic shapes.

There appear to be two outliers, the bulk of the shows cluster around 7, and a cluster of shows have ratings less than 3.5. During the *Seinfeld* week, except for the outliers, the shows were more uniformly distributed.

About 7 million people watched a typical show during the regular week, which is about 3 million people less than during the *Seinfeld* week. During the *Seinfeld* week, not only were there three shows watched by huge numbers of people, but also, in general, more people watched the programs. Even excluding the outliers, the spread for the regular week was less than for the *Seinfeld* week.

**E11. a.** ABC is approximately symmetrical and mound-shaped. CBS is slightly skewed right. FOX is more strongly skewed right. NBC is rather rectangular except for three outliers. UPN has very few shows (and very few viewers), and all but one are stacked on one point. WB is slightly skewed right. There are no outliers with the exception of NBC. There are no clusters or gaps.

**b.** The median for FOX is around 7 or 8. For NBC, the median is higher at about 12 or 13. The quartiles for FOX are about 6 and 14, and for NBC, they are about 9 and 16. So the middle half of the ratings have about the same spread for the two networks. The ratings are centered highest for NBC and lowest for UPN.

**c.** NBC has the most variability, and UPN has the least.

**d.** Answers will vary. A reasonable ranking would be NBC, CBS, ABC, FOX, WB, and UPN, based on their centers.

## 2.2 Graphical Displays for Distributions

---

### Objectives

- to learn the difference between a case and a variable and between quantitative and categorical variables
- to make and interpret the most common graphical displays: histogram, relative frequency histogram, stemplot (stem-and-leaf plot), bar graph, and dot plot

There should be a shared emphasis between making and interpreting graphical displays. On the AP Statistics exam, interpretation of plots is crucial, but occasionally students are asked to make a plot as well as interpret it. They must remember to include scales and labels, not just copy the “bare” plot from their calculator.

### Important Terms and Concepts

- case, quantitative variable, categorical variable
- histogram
- relative frequency and relative frequency histogram
- stemplot
- bar graph

### Lesson Planning

#### Class Time

One to two days. Again, we recommend moving quickly through this section. If your students are taking the Advanced Placement exam, you may wish to save some of the recommended practice problems and exercises to use for review right before the exam.

#### Materials

For Activity 2.2, a yardstick or meterstick

## Suggested Assignments

Classwork		
Essential	Recommended	Optional
D8, D12–D15, D17	D9, D10, D16	Activity 2.2
P9–P11, P13–P15	P8, P12, P16	D11

Homework		
Essential	Recommended	Optional
E12, E16, E20	E13, E14, E19, E22	E15, E17, E18, E21

### Lesson Notes: Variables and More About Dot Plots

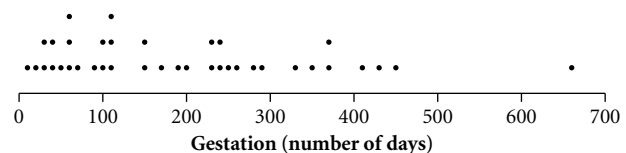
In this lesson, continue to emphasize that distributions, however they are displayed, should be described in terms of center, shape, and spread, and all plots should have scales on the axes. Popular graphing calculators do not include a dot plot option in their plot menus. However, a program can be loaded into your calculator. A dot plot program is found in the calculator guide. Boxplots appear in Section 2.3.

#### Discussion

- D8.** Quantitative variables are gestation period (in days), average longevity (in years), maximum longevity (in years), and speed (in mph). Categorical variables are whether a mammal is considered wild or not and whether a mammal is considered to be a predator or not.
- D9.**
- Of the 18 mammals for which speeds are given, 12 have speeds that end in 0 or 5.
  - Two-tenths of the 18 mammals, or 3.6
  - The most likely explanation is that the speeds are actually estimates for the wild mammals. Who is going to measure the speed of a grizzly bear in the wild? The speeds that don't end in 0 or 5 are for the dog, fox, giraffe, horse, pig, and squirrel. For these mammals, with the possible exception of the giraffe, you can see how speed could be measured accurately. (And it certainly is in horse races and dog races.)

#### Practice

- P8.** Quantitative variables are year of birth, year of hire, RIF stage, and age. Categorical variables are row number, job title, and pay category (hourly or salaried). Month of birth and month of hire fall somewhat in between and are best called “ordered categories.” Months are ordered and can be represented by numbers 1, 2, 3, . . . , 12 that can be meaningfully compared—they tell you whose birthdays come earlier in the year, for example. On the other hand, no one ever computes the mean or standard deviation of birth months as they might with age.
- P9.** The distribution is skewed right with no obvious gaps or clusters. There is a wall at 0 days because no mammal can have a smaller gestation period. The elephant is the only outlier. About half of the mammals have a gestation period of more than 160 days, and half have a shorter period. The middle half have gestation periods between 63 and 284. Large mammals have the longer gestation periods.



## Lesson Notes: Histograms

Graphing calculators generally put a value that falls on a boundary into the bar on the right. If you plot a histogram on a TI-83, for example, and press **TRACE**, you will see that the interval always includes the left endpoint and never the right endpoint.

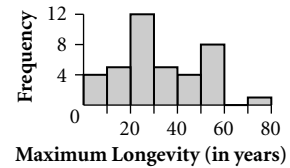
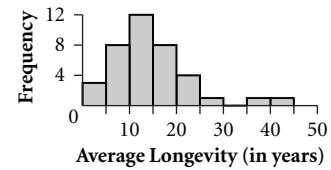
### Discussion

- D10.** The histograms are somewhat mound-shaped. In each histogram, the mammal speeds center around 40 miles per hour with a standard deviation of about 20 miles per hour. The bar width doesn't make a lot of difference for this data set, except that in the first histogram there is some hint of two peaks rather than one.
- D11.** The narrower bars cover a smaller interval on the real number line. Thus, you can state more precisely which speeds are in a given bar than you can when they are wider. If you made all histograms with very narrow bars, they would essentially be dot plots and you could have hundreds or thousands of bars. In a histogram, you combine nearby values into bars so that you can have fewer bars, making the overall shape easier to see. On the other hand, if the bars are too wide, you may miss gaps and clusters.
- D12.** No, the "skyline" of the histogram remains the same. Only the scale on the vertical axis changes. The histogram has a vertical scale from 0 to some integer, whereas the relative frequency histogram has a vertical scale from 0 to 1. From a relative frequency histogram, you cannot tell how many cases there are in each bar. From a frequency histogram, it is harder to judge relative frequency.

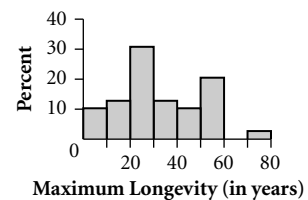
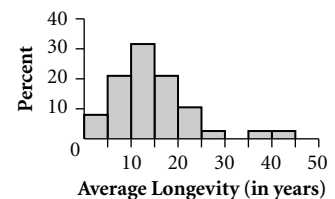
### Practice

- P10.** The shape of the maximum longevity distribution is quite different from that for average longevity. The average longevity distribution is skewed right with two possible outliers at 35–40 and 40–45. The distribution of maximum longevity is more uniform but with a peak at 20–30 years and an outlier at 70–80 years. As must be the case, the center of the distribution of maximum longevity is much higher than the center of the distribution of average longevity—about 30 years compared to about 15 years. The

spread of the distribution of maximum longevity is also much larger.



- P11.** The shapes do not change.



- P12.** The value at which the histogram balances (the mean) will be smaller than the value that divides the area into two equal parts (the median). If the distribution were symmetric, the mean and the median would be the same. However, here the outlying values tend to decrease the mean, but they don't affect the median. Specifically, if the values in the interval 40–48 were moved up to lie in the interval, say, 52–60, the median would be unchanged, but the mean would increase.

## Lesson Notes: Stemplots

John Tukey (United States, 1915–2000), a statistician at Bell Labs, invented the *stemplot* about 30 years ago. He also invented the boxplot and other useful techniques for exploratory data analysis. Because these techniques were invented so recently and are meant to be used for informal exploration, you will find variation from source to source about how they are constructed.

An efficient way to make a stemplot is to record the data as it appears in the list to create unordered leaves. Once all of the data have been recorded, redo the plot, ordering the leaves.

For more information on stemplots, consult *Exploring Data* by James M. Landwehr and Ann E. Watkins (D. Seymour Publications, 1995).

If your students will be taking the AP exam, it is also useful to discuss the Minitab stemplot in Display 2.32 in the student text on page 46 and to note that graphing calculators do not have the capability to make this plot without a specific program.

You may wish to have students display a histogram and a stemplot of the same data set and to compare and contrast the two displays, listing advantages and deficiencies (if any) of each plot.

### Discussion

**D13.** The stem-and-leaf plot shows a mound-shaped distribution in the middle with gaps at either end; the 11 and 12 at the low end and the 70 at the high end may be outliers. The median is 37, and the middle half of the values fall between 30 and 42. The stemplot gives more detail than the histogram.

**D14.** The leftmost column gives the number of values in the stemplot up to and including that row. Below the median, the counting is done from the bottom up. For example, the 8 at the beginning of the seventh row means that by the end of that row, with value 42, there are 8 values so far in the stemplot, counting from the bottom row up.

**D15.** This plot was made by Minitab statistical software. “Leaf Unit = 10” means that the leaf gives the tens place of the number and the stem gives the hundreds place. So, for example, the last line represents one number that falls in the interval 660 to 669.

Stem-and-leaf of Gestation N = 38  
Leaf Unit = 10 N\* = 1

```

6 0 123334
13 0 5666699
17 1 0001
(4) 1 5568
17 2 02334
12 2 5588
8 3 3
7 3 566
4 4 02
2 4 5
1 5
1 5
1 6
1 6 6

```

6|6 represents a number in the interval 660-669 days

### Practice

**P13.** The plots are shown here:

Stem-and-leaf of Ave Long N = 38  
Leaf Unit = 1.0 N\* = 1

```

3 0 134
11 0 55567788
(12) 1 000222222222
15 1 55555556
7 2 0000
3 2 5
2 3
2 3 5
1 4 1

```

4|1 represents 41 years

Stem-and-leaf of Max Long N = 39  
Leaf Unit = 1.0

```

1 0 4
4 0 588
8 1 3344
9 1 8
16 2 0000334
(5) 2 67778
18 3 0004
14 3 7
13 4 0
12 4 557
9 5 00000344
1 5
1 6
1 6
7 0 0

```

7|0 represents 70 years

Back-to-back stem-and-leaf plot

Ave Long	Max Long
431	0 4
88776555	0 588
222222222000	1 3344
65555555	1 8
0000	2 0000334
5	2 67778
	3 0004
5	3 7
1	4 0
	4 557
	5 00000344
	5
	6
	6
	7 0

5|2|6 represents an animal with an average life span of 25 years and an animal (not necessarily the same) with a maximum life span of 26 years.

As also seen in the histograms students made in P10, the values of average longevity are generally smaller and have a slight skewness toward the larger values. The distribution of maximum longevity is more spread out and is more uniform in shape.

Note, again, that there is evidence of estimating to the nearest 5 or 10. Of the 38 values of average longevity, half end in 5 or 0. Of the 39 values for maximum longevity, 17 end in 5 or 0.

**P14.** There aren't enough values to get a good idea of the shapes of the two distributions. However, it appears that there is an outlier on the high side for the predators and two outliers on the low side for the nonpredators. The median of the predator distribution, 40.5, is larger than for the nonpredators, 33.5. The spreads are about the same. The most striking thing about the two distributions is that there are no slow predators. That certainly makes sense because a slow predator wouldn't catch much prey.

**Activity 2.2: Do Units of Measurement Affect Your Estimates?**

This activity is optional. It has several purposes. The first is to give students practice in choosing an “appropriate and meaningful way to display data.” Use this activity and similar ones as an opportunity to discuss with students what makes a display appropriate. It is very important that students start to learn how to make good decisions based on the data they are dealing with.

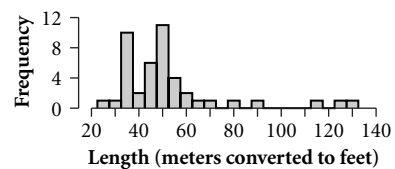
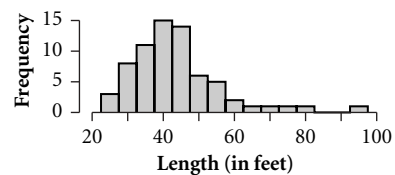
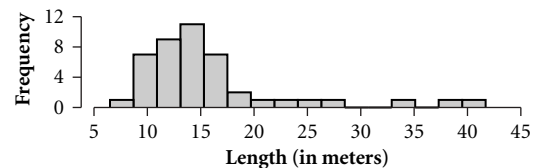
This activity also provides the opportunity to begin the discussion of the importance of random assignment in any experiment. Finally, you can use the activity to begin discussing the concept and the possibility of bias in measurement. That is, some methods of measurement are more likely to result in estimates that are too high (or too low) compared to other methods.

1. There will be several occasions when you must split your class randomly. The purpose of randomization is to create groups that are as similar as possible with the exception of the variable of interest. Here are some ways to split the class randomly.

- a. Have a slip of paper for each student in the class—half marked “feet” and half marked “meters.” Mix them up, and have each student draw a slip.
- b. Have students count off, and then randomly assign the evens to one group and the odds to the other. Decide by flipping a coin which group uses feet and which uses meters. Although this method is quick to use, it does not ensure a random division because there may be some pattern in the way students are seated (such as boy, girl, boy, . . .).
- c. Another way to split the class quickly is to have each student flip a coin, and those with heads estimate in feet while those with tails

estimate in meters. This is random, but you may not end up with groups approximately equal in size.

- 2. Watch to be sure that students who measure in meters do not estimate in feet and then convert to meters.
- 3. To compare the two sets of estimates, you should convert the estimates in meters to feet or vice versa. The formula for converting meters,  $m$ , to feet,  $f$ , is  $f = 0.305m$ . The formula for converting feet to meters is  $m = 3.28f$ .
- 4. An interesting set of data to compare your results with is given in Hand, et al., *Small Data Sets* (Chapman and Hall, 1994), p.2, where a similar experiment was done in Australia. For those students, meter was the familiar unit, of course. Histograms of the two groups are shown below. The length of the room in question was 13.1 meters, or 43 feet. Both of these distributions have centers remarkably close to the true value, and both are similarly skewed. Using 1 meter = 3.28 feet, the third histogram shows the meter data rescaled to feet. Now you can see that estimating in meters produced much more variation than estimating in feet.



5. It's possible that there are some differences between students on the two sides of the room. For example, students on the right side of the room might have a better view of the length than students on the left side, or perhaps students on one side of the room have to walk farther from the door to their desks and consequently have a better sense of

the size of the room. When there is any possibility of bias like this—and there almost always is—it’s best to randomize.

## Lesson Notes: Bar Graphs for Categorical Data

Quantitative variables are either continuous or discrete. If the numerical data can take on any value on a given interval, the variable is *continuous* and the data are called *measurement data*. Examples of continuous variables are height, weight, inflation rate in countries of the world, and blood pressure.

If the numerical data cannot take on every value within its range, the variable is *discrete*, and the data are also called discrete. Examples of discrete data are the number of cylinders in a car (which can only take on whole number values) and the sizes of a collection of wrenches (which can only take on values that are multiples of  $\frac{1}{16}$  inch).

Data that are actually discrete are often treated as continuous. For example, when you measure a person’s height, you must do so to, say, the nearest centimeter. That makes the data discrete, although the underlying distribution is continuous. Because they are always whole numbers, test scores such as the SAT or ACT are discrete but are often treated as measurement data. Looked at this way, there are no continuous real-world variables because the limitations of measuring instruments make all variables discrete.

Many statisticians say that a histogram should be used only for measurement data, not for discrete data. However, histograms are often suitable for discrete data if there are many values, as with ages or household income.

For bar graphs, the order of the bars, technically, is irrelevant. However, in many situations, one particular ordering might make more sense than other orderings. Sometimes alphabetical ordering of the categories is most reasonable, as when the categories are the 50 states. Sometimes ordering by height of the bar is most reasonable. Sometimes the categories have a natural ordering, as in Display 2.34 in the student text. See also D16, discussed next.

Categories can be coded using integers such as 0 or 1, as with predator/nonpredator and domestic/wild mammals, but these numbers are used merely for convenience and are not analyzed as measurements. The data are still categorical data.

## Discussion

**D16.** The ordering of the bars in Display 2.35 is completely arbitrary and could have been done in the opposite order. The categories in the education data in Display 2.34, however, represent increasing amounts of education and should be kept in order to see the pattern in the frequencies. Two types of categorical variables are those that have ordered categories and those that have unordered categories. A non-numerical example of ordered categories is small, medium, and large, as in shirt sizes.

- D17.**
- a.** The heights are the number of mammals in Display 2.24 that fall into that category. For example, the first bar shows that there are about 8 nonpredators that are domestic.
  - b.** Looking at the middle set of bars, for predators, the second bar is taller than the first. Thus, a predator is more likely to be wild than domestic.
  - c.** Looking at the first set of bars, for nonpredators, you can see that a nonpredator is also more likely to be wild because the second bar is taller than the first bar. However, for nonpredators, the first bar is a larger fraction of the second bar than is the case for predators. Thus, a predator is more likely to be wild than is a nonpredator.

*Note:* The way the bar graph is set up makes it easy to make the comparison asked for in part b but difficult to make the comparison in part c. You may wish to ask students to make a bar graph that makes it easy to answer part c. A two-way table like the one shown here can be helpful in summarizing the data on different categorical variables before making the bar graph.

	Nonpredator (0)	Predator (1)	Totals
Domestic (0)	8	2	10
Wild (1)	19	10	29
Totals	27	12	39

## Practice

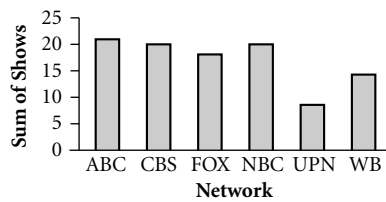
- P15.** The cases are the individual male members of the labor force aged 25 and older, and the variable is their educational attainment. The distribution shows an increasing proportion of males through the first four levels of education with a huge jump at the high school graduation level. After high school,

the proportions in each education category decrease regularly with increasing education levels, except for a spike at bachelor's degree.

The distributions for males and females have much the same shape, but females have lower proportions in categories 1 and 2 and higher proportions in all the other categories except 6, 8, and 9. The female labor force overall is a bit better educated than is the male labor force.

Relative frequency bar graphs are better for this comparison because the number of males and the number of females in the labor force are different.

**P16.**



### Exercises

- E12.** A case is a student in your class. The quantitative variables are age, number of siblings, and number of miles he or she lives from school. The categorical variables are hair color and gender.
- E13.** E13 could be done as a class discussion or a class activity if you have students collect their own pennies.
- These data are the ages of a set of pennies collected by a statistics class. A case is a penny. The variable is the age of the penny.
  - The shape is strongly skewed right. (In fact, the shape is characteristic of a geometric distribution, which students will study in Chapter 7.) The median is 8 years, and the spread is quite large, with the middle half of the ages of pennies falling between 3 and 15 years. However, it is not terribly unusual to see a penny that is more than 30 years old. (Typically, students predict that the shape will be normal.)
  - There are about the same number of pennies produced each year. Supposing that a penny has the same chance of going out of circulation each year, you would get a shape something like this one, with the height of each vertical line being a certain percentage of the previous one. The age of 0 is lower

than the age of 1 because the data were collected partway through the current year.

- E14. a.** Answers will vary depending on whether students think domestic or wild mammals live longer and what they think about the variability in the two classes.

**b.**

Domestic	Wild
4 0	13
85 0	556778
22220 1	0022222
5 1	5555556
0 2	000
2 5	
3	
3 5	
4 1	
1 5	represents 15 years

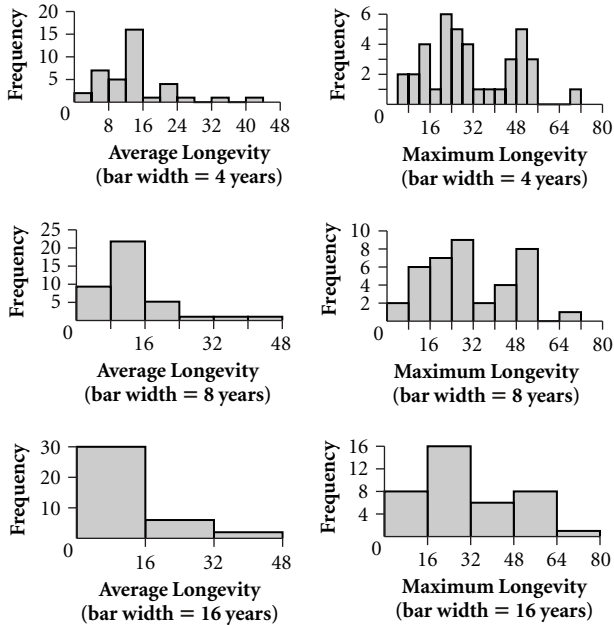
**c.** Both distributions appear to be slightly skewed to the right with possible outliers on the high side. The median of both distributions is 12, but the spread of the distribution for wild mammals is quite a bit larger. The middle half of the domestic mammals have an average longevity approximately between 8 and 12 years. The middle half of the wild mammals have an average longevity between 7.5 and 15.5 years.

- E15.** These are bar graphs. Answers will vary. In general, between 1992 and 1996, the number of outlets of five major fast-food chains in the United States grew by about 20% to 30%, but the average revenue stayed about the same or even went down.
- E16.**
- Graph D, because on an easy test most people get high scores.
  - Graph A, because the distribution of heights has two modes (mothers and daughters).
  - Graph C, because most countries in the Olympics get no medals at all and only a very small number of countries get multiple medals.
  - Graph B, because the weights should be mound-shaped. Most chickens will be clustered near a central weight with decreasing numbers having lower or much higher weights.
- E17.** The three histograms for average longevity are shown next with bar widths of 4, 8, and 16 years, respectively. At a bar width of 4, you can see a somewhat mound-shaped distribution with some skewness to the right, a pattern which is preserved at a bar width of 8.

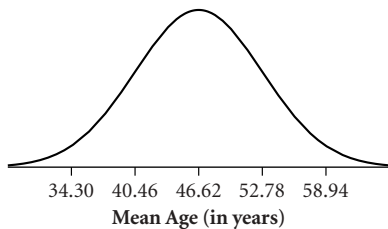
At a bar width of 16, the mound-shaped part disappears and the distribution only looks highly skewed.

The histograms for maximum longevity show a bimodal distribution at a bar width of 4; the bimodality is mostly obscured by the time the bar width gets to 16.

For both sets of data, a bar width of 16 is clearly too large, but both 4 and 8 work well. Note that in the stemplot in the answer to P13, the bar width is essentially 5.

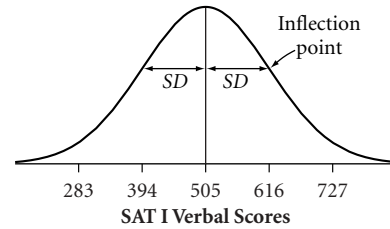


- E18. a.** The simulated means themselves have a mean of 46.62 years and a standard deviation of 6.16 years. Student estimates will differ somewhat. (The normal density curve that most closely corresponds to the histogram is shown here.)

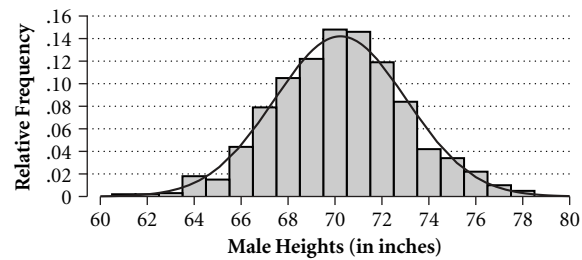


- b.** About 67%, about 95%; 100%. However, student estimates will vary, depending on their estimates of the mean and standard deviation.  
**c.** From the histogram, it appears that about 50 of the 1000 sample means, or 5%, had an average age of 58 or more.

- E19. a.** For the SAT I Math scores in 1999–2000, the mean was 514 and the standard deviation was 113. Students may estimate that the mean is about 500 because that is near the center of the distribution and that the standard deviation is about 100 because about two-thirds of the values lie between 400 and 600.  
**b.** About 67%; about 95%; 100%  
**c.** Answers will vary based on student estimates in part a.



- E20. a.**



- b.** The mean is 70.2 inches, and the standard deviation is 2.8 inches. Estimates from students should be within 1 inch of those.  
**c.** About .93, although student estimates will probably be a few percentage points different (about  $\pm .05$ ). Note that it is more efficient to find the proportion above 74 inches and then subtract from 1 or  $1 - (.005 + .01 + .022 + .034)$ .  
**d.** .163 or  $(.002 + .002 + .003 + .018 + .015 + .044 + .079)$ . Students' estimates will probably be a bit different ( $\pm .12$ ).  
**e.** Because the width of each bar is 1, the height and the area are the same. This is not true of every histogram. For example, see Display 2.41, which shows SAT I Math scores.  
**f.** A normal distribution is a smooth continuous curve that has a domain of  $-\infty$  to  $+\infty$ . The histogram of heights consists of distinct bars with flat tops and has a domain of only 60.5 to 78.5.

- E21.** For the United States, there is a population bulge around the ages of 30 to 50 for both men and women with decreasing percentages in the age groups above 50. For Mexico, the largest segments of the population are young children, with a regularly decreasing pattern in percentage of the population as the age increases. In both countries, there are more infant boys than infant girls. However, this reverses at the oldest ages, especially in the United States, where there are far more older women than older men.
- E22.** Students may find examples of bar graphs where the scale does not start at 0, making differences in the bars appear larger than they

actually are. They may find examples of bar graphs where the scale does not start at 0 and thus obscures differences that are important in the heights of the bars. They may find examples of “picture” graphs where a three-dimensional picture makes, for example, one quantity that is twice as big as another look like it is 8 times as big, and so on. The classic book, *How to Lie with Statistics* by Darrell Huff (Norton, 1993), originally published in 1965, is still informative and entertaining reading. More up-to-date examples can be found in Edward R. Tufte, *The Visual Display of Quantitative Information* (Graphics Press, 1983).

## 2.3 Measures of Center and Spread

---

### Objectives

- to compute and interpret the mean and median
- to compute the five-number summary and the interquartile range and to identify outliers
- to make and interpret boxplots
- to interpret percentiles and read cumulative relative frequency plots
- to compute and interpret the standard deviation
- to compute summary statistics from a frequency table
- to learn the effect on summary statistics of a linear transformation of the data
- to learn which summary statistics are resistant to outliers

In this section, students study the common measures of center (mean and median) and the common measures of spread (interquartile range and standard deviation). One of the main goals of the section is to help students realize that every measure of center should be accompanied by a corresponding measure of spread. A related goal is to convince students that the standard deviation is a reasonable measure of variation from the mean for data that are approximately normal.

### Important Terms and Concepts

- *measures of center* or *averages*: mean, median
- *measures of spread*: interquartile range (*IQR*), standard deviation, variance
- *five-number summary*: quartiles, minimum, maximum
- boxplot and modified boxplot
- percentiles
- cumulative frequency plots
- *linear transformations*: recentering and rescaling
- *outliers*: resistant to v. sensitive to
- frequency table
- formulas for the mean and standard deviation of a frequency table

### Lesson Planning

#### Class Time

Five to six days

#### Materials

For Activity 2.3, a ruler for each group of four students

## Suggested Assignments

Classwork		
Essential	Recommended	Optional
D18, D20, D22–D30, D33–D36 P17, P19–P21, P23–P25, P27, P29, P30, P32, P34, P35	Activity 2.3 D19, D31, D37 P18, P22, P28, P31, P33, P36	D21, D32 P26

Homework		
Essential	Recommended	Optional
E24, E26–E29, E33, E35, E40	E23, E25, E30, E37, E38, E41–E43	E31, E32, E34, E36, E39, E44

### Lesson Notes: Measures of Center

The symbol  $\bar{x}$  may be new to many of your students. Summation notation is introduced gradually in this text, and you probably will not need additional work with it.

#### Discussion

- D18.** a. mean: 2      median: 2  
 b. mean: 3      median: 3  
 c. mean: 4      median: 2  
 d. mean: 100      median: 2

The median is unchanged because increasing the largest number doesn't change the fact that 2 is the number in the center. The mean gets larger when any number is increased.

- D19.** a. As in D18, the mean is more affected by an outlier. In order to be the balance point, the mean has to move upward with the largest number because the mean increases if any number increases. In order to be the value in the center, the median doesn't have to change at all.  
 b. The distribution for the predators is skewed right, so the median is smaller than the mean. The distribution for the nonpredators is skewed left, so the median is larger than the mean. Because the nonpredators have generally smaller values to begin with, the means are farther apart than the medians.

c. There is a fairly large gap in the distribution between ages 38 and 48. When the larger values were removed, the central value or median had to “jump” that gap and became much smaller. This illustrates that the median can be quite unstable when there are only a few values in a distribution or when there are gaps.

#### Practice

- P17.** a. mean: 2.5      median: 2.5  
 b. mean: 3      median: 3  
 c. mean: 3.5      median: 3.5  
 d. mean: 49.5      median: 49.5  
 e. mean: 50      median: 50

**P18.** The mean height will increase by about 4 inches. The median should not change much because it will still be one of the 3rd graders, who all are about 4 feet tall.

**P19.** a. The measures of center for the life expectancies are

	Mean	Median
Africa	53.59	5
Europe	73.61	73

b. For Africa, the median is smaller than the mean because of the skewness toward the larger values. For Europe, the mean is about the same as the median.

## Lesson Notes: Measuring Spread Around the Median: Quartiles and IQR

A good way to show the need to connect a measure of spread with a measure of center is with this example: Two AP Statistics classes took the same test, and the median score for both classes was found to be 78. Can we conclude that the classes performed in the same way given only that their medians are equal? Two dot plots with the same center, 78, but very different spreads should convince the students of the need for more information than the center provides. For example,

Class 1: 78, 78, 78, 78, 78, 78, 78, 78, 78, 78

Class 2: 60 62 64 72 76 | 80 82 90 91 92

Students may have been introduced to quartiles previously. If not, be sure to use many visual displays such as the ones on page 58 of the student text to assist them in seeing the idea.

A common mistake students make is to divide the range into four equal parts. For example, with the data 60, 62, 64, 72, 76, 80, 82, 90, 91, and 92, a student may divide the range of 32 into fourths, or 8, and get

$$\text{Min} = 60$$

$$Q_1 = 68$$

$$Q_2 = 76$$

$$Q_3 = 84$$

$$\text{Max} = 92$$

Note that the correct quartiles are 64, 78, and 90.

On the AP Statistics exam, students are expected to understand, for example, that the first quartile is the number that divides the first and second quarters. It is incorrect to say something like, “My test score was really low. It fell in the first quartile.” It would be correct to say, “It fell below the first quartile” or “It fell in the lowest quarter of all scores.”

### Discussion

**D20. a.** The middle half of the speeds of domestic mammals are between 30 and 40. The spread for wild mammals is a bit larger—the middle half of the speeds are between 27.5 and 43.5. Half of the domestic mammals have speeds above 37 and half below. The median for wild mammals, again, is almost the same, at 36. Note that in each case the median is

closer to the upper quartile than the lower quartile, indicating that the distribution of speeds may be skewed left.

**b.** The wild mammals are more likely to be predators than are the domestic animals (see D17), and the speeds of predators have the larger IQR (see Display 2.31).

**D21.** Detective Seymour has received “quite a few” descriptions of the suspect from various eyewitnesses. These descriptions varied so much that Detective Seymour felt that the average was useless. For example, suppose he had four eyewitnesses, two of whom said the murderer was 5’7” tall and two of whom said the murderer was 5’8” tall. In this case, it would be perfectly reasonable for Detective Seymour to believe that the murderer was close to 5’7½” tall. However, if his four eyewitnesses said the murderer was 5’, 5’4”, 5’9”, and 6’5”, it wouldn’t be reasonable for him to make any conclusion about the murderer’s height even though the average is still 5’7½”.

### Practice

**P20. a.** quartiles: 2 and 5      IQR: 3

**b.** quartiles: 2 and 6      IQR: 4

**c.** quartiles: 2.5 and 6.5      IQR: 4

**d.** quartiles: 2.5 and 7.5      IQR: 5

**P21. a.** The quartiles and medians are marked in bold on this plot.

Predators	Nonpredators
1	0 34
75	0 556788
22222	1 0002222
65	1 555555
	2 0000
5	2
	3
	3 5
	4 1

1|5 represents 15 years

**b.** The distribution of the average longevity of predators is mound-shaped, centered at about 12 years, with 50% of the values falling between 7 and 15 years. The distribution of the average longevity of nonpredators is centered at exactly the same place and has about the same spread, but it has two outliers on the high side.

## Lesson Notes: Five-Number Summaries, Outliers, and Boxplots

### Quartiles

Some textbooks and software packages use rules for finding quartiles that differ somewhat from that in *Statistics in Action*. For example, when there is an odd number of values, the median may be included in each half when computing the quartiles. Some software packages (such as Minitab and DataDesk) use the formula  $\frac{(n+1)}{4}$  for the position of the first quartile. Texas Instruments calculators use the rule given in *Statistics in Action*.

It is not important for students to understand all of these various rules. They need to know the basic idea that the quartiles, with the median, divide the data into four parts with roughly the same number of values.

For large sets of data, the rule that is used makes little difference. For smaller sets of data, especially those with gaps, it can make quite a difference. For those types of data sets, a five-number summary is inappropriate, as in this example:

Consider the data set of the number of hours of TV per week a group of 8th graders said they watched on Monday through Friday nights during the school year: {1, 1, 2, 8, 9, 10, 19, 20}.

Two five-number summaries can be calculated:

<i>Statistics in Action</i> (TI-83)	Minitab
Min = 1	Min = 1
Q1 = 1.5	Q1 = 1.25
Med = 8.5	Med = 8.5
Q3 = 14.5	Q3 = 16.75
Max = 20	Max = 20

### Outliers

Why is the number 1.5 used in the definition of outlier? According to John Tukey, who defined outliers this way, the reason is that 1 is too small and 2 is too large. That is, if you use the value 1, you get too many values defined as outliers that don't seem to be outliers. If you use 2, values that you think should be outliers aren't defined as outliers. Although the 1.5 rule is somewhat arbitrary, it works well in practice due to the remarkable intuition of an expert data analyst. There are methods of defining outliers other than the  $1.5 \cdot IQR$  rule, but the  $1.5 \cdot IQR$  method is a convention that many statisticians employ.

Different software packages use different symbols to mark outliers, and some mark outliers more than  $3 \cdot IQR$  from the quartiles differently from those only  $1.5 \cdot IQR$  from the quartiles. In addition, some software packages and many calculators orient the boxes horizontally.

Boxplots display skewness and show the existence of outliers in distributions. This fact makes it the most useful plot in Chapter 9 for deciding whether the conditions for the  $t$ -test have been met. Once the students have computed the five-number summary and constructed both regular and modified boxplots for several examples, students who will be taking the AP Statistics exam should know how to do this on their calculators and should be able to read the print-outs of common statistical software.

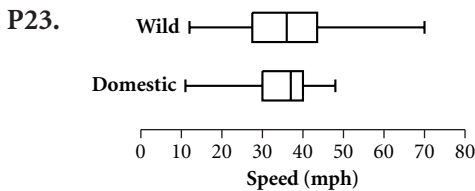
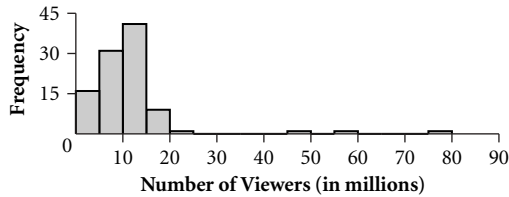
### Discussion

- D22.** It gives the value of the quartiles, not their position; the median.
- D23.** a. 50%; 25%; 25%
- b. A boxplot for a data set that is extremely skewed right would have a lower whisker that is shorter than the bottom half of the box, which is shorter than the top half of the box, which is shorter than the upper whisker. A boxplot for a data set that is extremely skewed left would have a lower whisker that is longer than the lower half of the box, which is longer than the upper half of the box, which is longer than the upper whisker. A boxplot for a data set that is symmetric would have whiskers of equal length, and the two halves of the box would be of equal length.
- c. The  $IQR$  is equal to the length of the box. The range is the distance from the end of one whisker (or outermost outlier) to the end of the other whisker (or outermost outlier).
- d. From a boxplot, you can see the five-number summary exactly and outliers are clearly marked. These must be estimated from a histogram and can be difficult to estimate. From a histogram, you can estimate the mean by estimating a balance point for the distribution. You cannot do this with a boxplot. A histogram will reveal the frequency of the data within an interval. You do not know the exact values, but you know how many are within the given boundaries.

You know a lower and upper bound but not necessarily the exact least and greatest value. You know where there are clusters of data and where there are gaps. With a boxplot, you get a sense of the basic shape of the distribution, but you cannot see clusters or gaps.

**Practice**

- P22. a. *Seinfeld*, *Seinfeld Clips*, and *ER*  
 b. *Touched by an Angel*  
 c. The actual histogram is shown here:



- P24. a. The five-number summary for the average longevity of this set of mammals is
- minimum: 1
  - lower quartile ( $Q_1$ ): 8
  - median: 12
  - upper quartile ( $Q_3$ ): 15
  - maximum: 41

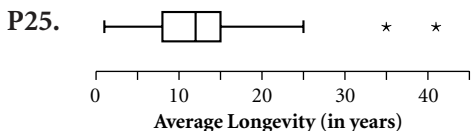
Once again, software packages may give quartiles slightly different from those done by hand. Here are the results from Minitab:

Variable	N	N*	Mean	Median	TrMean	StDev	SEMean
Ave long	38	1	13.13	12.00	12.32	8.00	1.30

Variable	Min	Max	Q1	Q3
Ave long	1.00	41.00	7.75	15.00

- b.  $IQR = 15 - 8 = 7$   
 c.  $Q_1 - 1.5 \cdot IQR = 8 - 1.5(7) = -2.5$ . There are no outliers on the lower end.  
 d.  $Q_3 + 1.5 \cdot IQR = 15 + 1.5(7) = 25.5$ . The life spans of 35 years for the elephant and 41 years for the hippopotamus are outliers. The largest value that isn't an outlier is 25. This is where the upper whisker will end.



- P26. Yes. There will be no lower whisker, for example, if the minimum and the first quartile are equal. This set of data has no lower whisker:  
 $\{1, 1, 1, 1, 2, 3, 5, 6, 7, 12, 14, 16\}$ .

**Lesson Notes: Percentiles and Cumulative Frequency Plots**

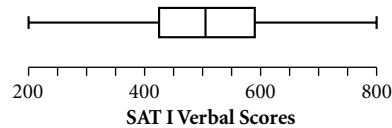
We say that a given score is at, say, the 74th percentile if 74% of the scores in the distribution lie at or below the given score.

**Discussion**

- D24. a. about the 23rd percentile  
 b. about 325 to 700; about 290 to 745  
 c. about 480  
 D25. 90%; between the 2.5th percentile and the 97.5th percentile

**Practice**

- P27. The quartiles are about 425 and 590; the median is about 505; the *IQR* is about  $590 - 425 = 165$ . The minimum of 200 and the maximum of 800 complete the five-number summary. A boxplot is shown here:



**Lesson Notes: Measuring Spread About the Mean: The Standard Deviation**

**Activity 2.3: Comparing Hand Spans: How Far Are You from the Mean?**

The activity “Comparing Hand Spans” familiarizes students with the concept of a measure of variation from the mean. They also learn that several such measures may be defined. Students should be in groups of four or five for this activity.

- 1–3. Measurements, means, and dot plots will vary.
4. The major source of variability is that everyone in the group has a different hand span, depending on the size of their hand and on their flexibility. The other source of variability is in the uncertainty of the measurement of a given hand. That is, if different people measure the same hand, they may

get different values. Also, if the same hand is measured at different times, the measurement for that hand span may be a little different each time.

5. Answers will vary.

6–8. Groups will tend to invent the mean absolute deviation (*MAD*):

$$MAD = \frac{\sum |x - \bar{x}|}{n}$$

The *MAD* is easily understood and a reasonable measure of spread for describing a distribution of data but has almost no importance in statistical theory. Sometimes a group will invent a statistic defined as the number of group members whose hand span is not equal to the mean (the number of “misses”). Other groups may suggest the largest difference from the mean.

### The Standard Deviation

Have students calculate the standard deviation by hand a few times (using a table for organization) so that the formula is more meaningful for them.

Two questions will surface when students see the formula for the standard deviation: Why do we square the deviations instead of taking the absolute value, and why and when do we divide by  $n$  and by  $n - 1$ ?

**Why do we square rather than take the absolute value?** In Activity 2.3, students may invent the mean absolute deviation (*MAD*):

$$MAD = \frac{\sum |x - \bar{x}|}{n}$$

as a method of measuring spread. Intuitively, this is a perfectly reasonable measure and much simpler than the standard deviation. Why, then, don't statisticians use it?

There are two main uses of statistics that we should try to keep separate when deciding what summary measures to use. The first is exploratory, in which we are just trying to describe key features of the data. The second is inferential, in which we are trying to infer something about a larger population (sample survey) or about treatments (experiments). In the exploratory sense, many measures of variability will work (interquartile range, mean absolute deviation, standard deviation, etc.), depending on the nature of the data. In fact, the standard deviation is not always useful for describing distributions of real data. It works well as a measure of spread only for data distributions that

are approximately normal in shape. If the distribution is skewed, the *IQR* is usually a better choice.

However, the standard deviation cannot be avoided (at least easily avoided) when one is using classical inference for means. The sampling distribution of the sample mean is approximately normal (given a reasonable sample size) with standard deviation equal to the population standard deviation divided by the square root of sample size. Thus, confidence intervals and test statistics for means require a good estimate of the population standard deviation, which is provided by the sample standard deviation.

In short, in exploratory analysis, we can do anything reasonable after viewing the shape of the data distribution. In inference, we use the sample standard deviation as an estimate of the population standard deviation because we are forced into the normal model by the laws of probability.

One reason that the standard deviation comes up so often in statistical theory is that squaring rather than taking the absolute value has an important characteristic. Have students examine the difference among the graphs of  $y = |x|$ ,  $y = |x| + |x - 2|$ ,  $y = |x| + |x - 2| + |x - 4|$ , and so on. Observe that the graphs are composed of segments and rays and the pattern varies depending on the number of terms. Then explore  $y = x^2$ ,  $y = x^2 + (x - 2)^2$ ,  $y = x^2 + (x - 2)^2 + (x - 4)^2$ , and so on. Observe that in each case the graph is a parabola. If the situation is to be analyzed mathematically, squaring is simpler than absolute values.

**Why do we divide by  $n - 1$  rather than by  $n$ ?**

Essentially, the reason for dividing by  $n - 1$  is to adjust for working with a sample. As mentioned earlier, the population standard deviation in inferential statistics is estimated by the sample standard deviation. The variability in a random sample tends to be less than in the entire population. Thus, you divide by  $n - 1$  rather than by  $n$  to increase the sample variability a bit. As the sample size increases, it makes little difference whether you divide by  $n$  or by  $n - 1$ . This issue will be investigated more fully in Chapter 5 along with the discussion of sampling distributions.

### Discussion

**D26.** Yes, this seems reasonable. Some values are more than this distance from the mean, and some are less. In addition, many of the deviations are close to the value of the standard deviation.

- D27.** Each of these, except the variance, has the same units as the data—*years*. The variance is in *years*<sup>2</sup>.
- D28.** Dividing by a slightly smaller number makes the standard deviation a bit larger.
- D29.** In both cases, the formulas involve the square root of a sum of squared differences. To compute the standard deviation, find the differences, square, average, and take the square root. To find the distance between two points, find the differences, square, add, and take the square root. For example, suppose that you have a set of three values, say, {1, 4, 10}, with mean 5. Then, except for dividing by the sample size, the standard deviation is the same thing as the distance in space between the point (1, 4, 10) and the point of the mean (5, 5, 5):

$$\sqrt{(1 - 5)^2 + (4 - 5)^2 + (10 - 5)^2}$$

Thus, the measure of distance in statistics, the standard deviation, has almost exactly the same form as the measure of distance in Euclidean geometry. Perhaps this reason, above all others, helps convince students that the standard deviation is a natural measure of spread.

### Practice

- P28.** The mean is 4.4, and the deviations from the mean are shown in this table. The sum of these deviations is 0. To get the standard deviation, find the  $(n - 1)$  average of the squared deviations,  $\frac{41.2}{4} = 10.3$ , and take the square root to get about 3.21.

Value $x$	Deviation from Mean: $x - \bar{x}$	Squared Deviations: $(x - \bar{x})^2$
1	-3.4	11.56
2	-2.4	5.76
4	-0.4	0.16
6	1.6	2.56
9	4.6	21.16
<b>Sum</b>	<b>0</b>	<b>41.20</b>

- P29.** a. i. 0, because all of the deviations from the mean are 0  
 b. iii. 0.577  
 c. iv. 1.581  
 d. vii. 5.774. Note that the values are 10 times as far from the mean as those in part b,

- so the standard deviation is 10 times as large.  
 e. ii. 0.058. Note that the values are one-tenth as far from the mean as those in part b, so the standard deviation is one-tenth as large.  
 f. v. 3.162. Note that the values are twice as far apart as those in part c, so the standard deviation is twice as large.  
 g. vi. 3.606. It may be hard for students to distinguish part f from part g. If so, they should compute the standard deviation to check their answer.

## Lesson Notes: Which Summary Statistic?

### Discussion

Plotting the data is the first step to making good decisions about summary statistics. Many students may ask for “rules.” There are very few hard and fast rules when it comes to data analysis.

- D30.** Multiply the number of houses by the mean value by the tax rate. The total is equal to the number of values times the mean. If you know the number of houses, you can easily convert between the mean value and the total value.
- D31.** Income is strongly skewed right. A few people make a lot of money, whereas most people are clustered together toward the low end. Consequently, the mean looks larger than most people think is “typical.” The median tells you that half of the residents earn more and half earn less. Another reason the median may be given is that the median income is probably easier to estimate.
- D32.** a. If you knew the upper quartile, you could make the seats wide enough for 75% of the people. The maximum wouldn’t do much good because it would be so large that it would be too expensive to make every seat that size. Perhaps the best thing would be to know (say) the 95th percentile, so you could make the seats wide enough for 95% of the people.  
 b. If you want the best price, you need to know the minimum.  
 c. If you tend to study less than most people, the lower quartile of the study time needed by students at that college would give you a good indication of how much time you

will need. If you are typical, you might like to know the median. If you tend to study more than most people, you might like to know the upper quartile.

### Practice

- P30.** The mean is larger than the median because house values tend to be skewed right—some houses cost a lot more than most houses in a community. Very few houses cost a lot less. To get the total property taxes, multiply the number of houses by the mean value by the tax rate to get  $(\$392,059)(9,751)(0.0115) = \$43,964,124$ . This is an average of \$4,508.68 per house. (This assumes that the assessed value is equal to the price. California's Proposition 13 makes the situation more complicated because many houses are assessed at quite a bit under the actual value.)
- P31.** **a.** Medians were used in this story because the distribution of car ages is strongly skewed right. There are more brand-new cars on the road than cars of any other age (because cars of any other age have been disappearing due to accidents and mechanical problems). A few people drive very old cars. A story that appeared in the same newspaper tells of such a couple—the husband drives a 1971 Blazer, and the wife drives a 1966 Mustang.
- b.** The article says that the year 1970 was “before the Big Three auto makers were challenged by a flood of well-built Japanese imports” and quotes an expert as saying vehicles are proving more durable. Another reasonable explanation that students might give is that people are choosing to spend their income on other things or that they are forced to spend their income on other things. Again, encourage students to generate lots of possibilities.

## Lesson Notes: The Effects of Recentering and Rescaling

### Discussion

- D33.** **a.** mean: 188 pesos; *SD*: 47 pesos  
**b.** median: \$10;  $Q_1$ : \$5;  $Q_3$ : \$20

### Practice

- P32.** **a.** Divide each of the summary statistics by 12 so then the mean is 4 feet, the median is 3.75 feet, the standard deviation is 0.2 feet, and the interquartile range is 0.25 feet.
- b.** The mean is 50 inches, and the median is 47 inches. The standard deviation and interquartile range do not change.
- c.** The mean is  $4\frac{1}{3}$  feet, the median is  $4\frac{1}{12}$  feet, the standard deviation is 0.2 feet, and the interquartile range is 0.25 feet.
- P33.** **a.** mean: 2; *SD*: 1  
**b.** mean: 12; *SD*: 1 (same as in part a)  
**c.** mean: 20; *SD*: 10 (10 times that in part a)  
**d.** mean: 110; *SD*: 5 (5 times that in part a)  
**e.** mean:  $-900$ ; *SD*: 100 (100 times that in part a)

## Lesson Notes: The Influence of Outliers

Although it is always true that removing a value that is larger than the rest of the values in a data set will decrease the mean, it does not necessarily decrease the standard deviation. The reason is that if the mean decreases, it changes all of the deviations from the mean. It's also possible to remove the only outlier in a set of data only to create a new outlier!

### Discussion

- D34.** **a.** The mean is affected by the outliers. Whether the effect is moderate or not depends on the context. To compute the mean, first add the values. An unusually large value will increase this sum quite a lot. Outliers tend to have greater influence in small samples than in large ones.
- b.** The numerical value of an outlier does not affect the median because the size of the largest or smallest value doesn't affect which value is in the center. But removal of an outlier from a data set may cause some change in the median because of a change in the sample size.
- D35.** **a.** Yes. The range is greatly affected by an outlier because it is computed by subtracting the largest and smallest values.

b. Yes. Generally, the standard deviation is affected greatly by an outlier. To compute the standard deviation, first square the differences from the mean. If one of these differences is large, squaring it makes it even larger.

c. No. An outlier does not affect the interquartile range because the quartiles aren't affected by the size of the maximum or the minimum.

### Practice

P34. a. Outliers occur above  $-30 + 1.5(21) = 1.5$ . Hawaii, at 12, is an outlier.

b. The count will decrease by 1 to become 49.

Summary of Lowest Temperature without Hawaii  
No selector

Percentile	25
Count	49
Mean	-41.5
Median	-40
StdDev	16.2
Min	-80
Max	-2
Range	78
Lower ith %tile	-51
Upper ith %tile	-32

The minimum, median, quartiles, and interquartile range should remain the same or about the same. With a sample of size 50, removing one data value will have little effect on the median but could have greater effect on the quartiles as they are, in essence, the medians of samples of size 25. The mean should go down by a bit more than one degree because the difference between Hawaii's temperature and the mean is about  $-52$  degrees and there are 49 states remaining. The standard deviation should go down only slightly, but this is difficult to predict. The range will decrease from 92 degrees to about 80 degrees. The maximum will decrease from 12 to a little less than zero.

## Lesson Notes: Summaries from a Frequency Table

The formulas in this section provide a foundation for calculating the mean and standard deviation of probability distributions.

On some calculators, such as the TI-83, you can put the values from a frequency table into, say, list L1. Then put the frequencies into list L2. Using the command 1-VAR STAT L1, L2 will give the summary statistics for the frequency table. Similarly, you can make a histogram.

### Discussion

D36. a. Skewed right, or toward the larger values. There is a wall at 0 because no family can have fewer than zero children.

b. Because there are 100 families, the median is at position  $\frac{(100 + 1)}{2} = 50.5$  between family 50 and 51. For 1967, the 50th and 51st families both occur in the category of 3 children, so 3 is the median.

c. For 1967, the computation for mean is shown below.

D37. The formula uses multiplication as a short cut for addition. Instead of adding the value  $x$  a total of  $f$  times, you can just multiply  $x$  by  $f$ . The symbol  $n$  stands for the sample size or 100,  $x$  is the value or number of children, and  $f$  is the frequency or number of families that have that particular number of children.

Again, the formula substitutes multiplication for repeated addition.

### Practice

P35. a. For 1997, the mean is shown at the bottom of page 49. You can find the standard deviation for 1997 by realizing that there are 15 deviations of  $(0 - 2.3)$ , 22 deviations of  $(1 - 2.3)$ , and so on. The standard deviation is also shown on page 49. The standard

### Lesson 2.3, D36c

$$\begin{aligned}\bar{x} &= \frac{0 \cdot 5 + 1 \cdot 10 + 2 \cdot 21 + 3 \cdot 28 + 4 \cdot 17 + 5 \cdot 7 + 6 \cdot 4 + 7 \cdot 3 + 8 \cdot 5}{100} \\ &= \frac{324}{100} \\ &= 3.24 \\ SD &= 1.89\end{aligned}$$

deviation for 1997 is 1.84, which is a bit less than that for 1967, which is 1.89.

**b.** For 1997, the 50th and 51st families both occur in the category of 2 children, so that is the median, one fewer than in 1967.

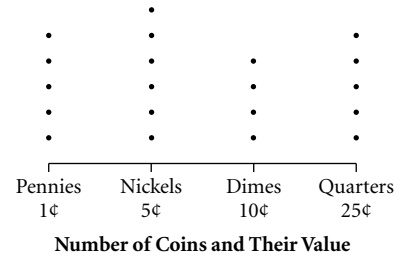
**c.** The quartiles are at positions 25.5 and 75.5. Counting in from the top of the 1967 table, the 25th and 26th families both occur in the category of 2 children, so that is the first or lower quartile. For 1967, the quartiles are 2 and 4. For 1997, the quartiles are 1 and 3. In both cases, the *IQR* is 2.

**d.** The median number of children for 1997 is 2, 1 fewer than in 1967; the mean also went down by about 1 child per family, from 3.2 to 2.3. The distributions kept the same shape and about the same spread.

*Note about outliers and skewed distributions:* Note that both the 1967 and 1997 distributions have values identified as possible outliers even though no value is separated from the rest of the distribution. For 1997, an outlier would be more than

$3 + 1.5(2) = 6$  or less than  $1 - 1.5(2) = -2$ . An outlier is a family with more than 6 children. For 1967, any outlier would have to lie below  $2 - 1.5(2) = -1$  or above  $4 + 1.5(2) = 7$ . So two families would be identified as outliers, those with 8 children. A distribution with long tails will usually have values identified as outliers even though they are not separated from the rest of the values.

**P36. a.** The mean is about 10.



- b.** The mean is shown below.  
**c.** The standard deviation is closest to 10.  
**d.** The standard deviation is shown below.

**Lesson 2.3, P35a**

Mean:

$$\begin{aligned}\bar{x} &= \frac{0 \cdot 15 + 1 \cdot 22 + 2 \cdot 25 + 3 \cdot 18 + 4 \cdot 10 + 5 \cdot 2 + 6 \cdot 4 + 7 \cdot 2 + 8 \cdot 2}{100} \\ &= \frac{230}{100} \\ &= 2.3\end{aligned}$$

Standard Deviation:

$$SD = \sqrt{\frac{(0 - 2.3)^2 \cdot 15 + (1 - 2.3)^2 \cdot 22 + \dots + (8 - 2.3)^2 \cdot 2}{99}} \approx 1.84$$

**Lesson 2.3, P36b**

$$\bar{x} = \frac{\sum x \cdot f}{n} = \frac{1 \cdot 5 + 5 \cdot 6 + 10 \cdot 4 + 25 \cdot 5}{20} = \frac{200}{20} = 10$$

**Lesson 2.3, P36d**

$$\begin{aligned}s &= \sqrt{\frac{\sum (x - \bar{x})^2 \cdot f}{n - 1}} \\ &= \sqrt{\frac{(1 - 10)^2 \cdot 5 + (5 - 10)^2 \cdot 6 + (10 - 10)^2 \cdot 4 + (25 - 10)^2 \cdot 5}{99}} \\ &\approx 9.4\end{aligned}$$

## Exercises

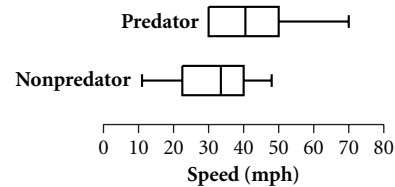
**E23. a.** Either could be used depending on the purpose of computing a measure of center. If, as is typical, there are a few expensive homes mixed in with many modestly priced homes, then the mean price will be larger than the median price. So real estate agents usually report the median price because it is lower and it tells people that half the prices are lower and half are higher. The tax collector would be interested in the mean price because the mean times the tax rate times the number of houses gives the total taxes collected.

**b.** As always, the answer depends on the purpose for computing a measure of center. Most likely, the reason here is to establish the total crop in Iowa for the year. In that case, it is best to find the mean yield per acre. This mean could be multiplied by the total acres planted in corn to approximate a total yield. An individual farmer probably would want to know the median as it gives the better indication of whether his or her yield was typical.

**c.** Again, the purpose of computing the measure of center determines which one you would use. Survival times are usually strongly skewed right. Telling a patient only the mean survival time would give too optimistic a picture. The smaller median would inform the person that half the people survive longer and half shorter. On the other hand, if you are the physician and must allocate your time by estimating the total number of hours you will be caring for your patients with this disease, the mean would be better. You would then multiply the mean number of survival days by the number of patients you have by an estimate of the number of hours each day that each patient takes.

- E24. a.** boxplot 3  
**b.** boxplot 1  
**c.** boxplot 2

**E25.** The back-to-back stemplot is better because there are only a few values, so you may as well see them.



**E26.** The third boxplot cannot be the plot for both classes combined because the minimum test score for the second period is about 10, and that would be the lowest for the combined set also.

**E27.** The seventh value is 10, as you can see from solving

$$25 = \frac{x + 24 + 47 + 34 + 10 + 22 + 28}{7}$$

**E28. a.** II has the largest standard deviation, and III has the smallest.

**b.** II and III

**E29.** The second data set is the same as the first except that each value is 4 more, so the spreads are the same. The standard deviation of the recentered values, then, is also about 30.

**E30.** The set of heights of all female NCAA basketball players will have the larger mean because basketball players tend to be taller than other athletes, in general. The set of all female NCAA athletes will have the larger standard deviation because it will include tall, medium, and short athletes, whereas the set of all basketball players will include mostly tall athletes.

**E31. a.** The mean length of a generation. You would divide 300 years by the mean length of a generation to get the number of generations.

**b.** The average speed. You multiply the average speed by the time to get the distance.

**c.** Yes, if you know the number of trees. The average volume is  $\pi r^2 h = \pi \cdot 3^2 \cdot 45 \approx 1272 \text{ ft}^3$ . To get the total volume, multiply by the number of trees.

**E32. a.** Replace the 15 with a 1.

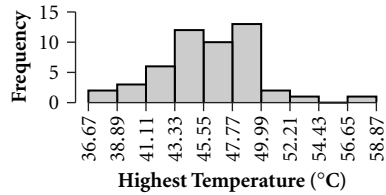
**b.** Replace the 32 with a 10.

**c.** There is no way to get an outlier.

- E33. a.** To make the histogram, students need only copy the histogram in the student text and then use the formula

$$C = \frac{5}{9}(F - 32)$$

to convert the numbers on the  $x$ -axis from °F to °C. The scale would then go from 36.67 to 58.89.



- b.** Note that the standard deviation is the tricky one: You just multiply by  $\frac{5}{9}$ . For each of the others, you subtract 32 and then multiply by  $\frac{5}{9}$ .

Variable	N	Mean	Median	TrMean	StDev
Highest	50	45.61	45.56	45.53	3.72
Variable	Min	Max	Q1	Q3	
Highest	37.78	56.67	43.33	47.78	

- c.** Yes, there is an outlier on the high side. The  $IQR = 4.45$ , and  $1.5(IQR) = 6.675$ . So  $Q_3 + 1.5 \cdot IQR = 54.455$ , and the maximum is larger than that—so definitely an outlier.  $Q_1 - 1.5 \cdot IQR = 36.655$ , and the minimum is bigger—so none on the lower end.

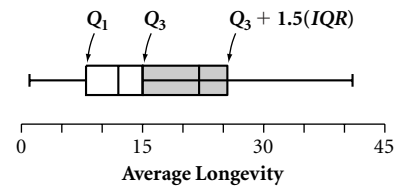
- E34. a.** i. 6.325 v. 6.667  
 ii. 2.000 v. 2.010  
 iii. 0.632 v. 0.633
- b.** No, as  $n$  gets larger, the difference between  $s$  and  $\sigma$  goes to 0.
- E35. a.** 3.11 grams  
**b.** 0.04 gram  
**c.** Yes, most of the weights are between 3.07 and 3.15.
- E36.** The mean is 73.154, and the standard deviation is 3.065.
- E37.** Of the 36 salaried workers, 18 were kept and 18 were laid off. The 18 salaried workers who were laid off had a median age of 53.5 and quartiles 42 and 61. The 18 salaried workers who were kept had a median age of 48 with quartiles 37 and 55. So the median age of the workers laid off was 5.5 years older, but the distributions had about the same  $IQR$ .

Salaried Worker's Ages

Kept	Laid Off
2	3
9	2
421	3 012
7	3
2	4 2
8887	4 9
443	5 0234
975	5 669
10	6 134
	6 169

6|9 represents 69 years

- E38.** The maximum value must be an outlier. The length of the box is the  $IQR$ , and  $Q_3$  lies at the top of the box. An outlier lies beyond  $Q_3 + 1.5(IQR)$ . This point can be estimated from the boxplot by imagining stacking one-and-a-half boxes to the right of the box as illustrated here. The top of these boxes is less than the maximum value.



- E39.** There are three outliers in Display 2.52 but only two in Display 2.74. The explanation is that the boxplot for the average longevity of domestic mammals creates its own outlier because the spread in longevity in these mammals is otherwise so small. There is no reason to expect that the outliers of subsets of a set of data will be the same as in the set of data itself.
- E40.** The standard deviation is 0 because the numbers don't vary.
- E41.** First, subtract 5478 from every number, leaving

0.1    0.3    0.3    0.9    0.4    0.2

The mean of these numbers is 0.3667, so the mean of the original numbers is 5478.3667. The standard deviation is about 0.280. Because recentering doesn't change the standard deviation, that's the standard deviation for the original set of numbers, too.

**E42.** Let the mean of the original set of data be

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

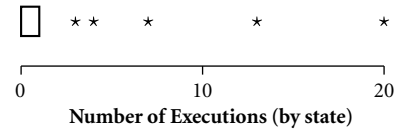
Then the mean of the transformed data is the result of the equation shown below.

**E43.** If  $\bar{x}$  is the mean of the original set of data, then the standard deviation is shown below.

You know from E42 that the mean of the transformed data is  $\bar{x} + c$ . Then the standard deviation of the transformed data is shown below.

**E44.** Mean: 1.36  
 Median: 0  
 First quartile: 0  
 Third quartile: 1

The *IQR* is 1, so the 8 values larger than  $1 + 1.5(1) = 2.5$  are outliers. This results in an odd-looking boxplot with no median line and no whiskers.



**Lesson 2.3, E42**

$$\begin{aligned} & \frac{(x_1 + c) + (x_2 + c) + (x_3 + c) + (x_4 + c) + (x_5 + c)}{5} \\ &= \frac{x_1 + x_2 + x_3 + x_4 + x_5 + 5c}{5} \\ &= \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} + \frac{5c}{5} \\ &= \bar{x} + c \end{aligned}$$

**Lesson 2.3, E43**

Standard Deviation of Original Data:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2}{4}}$$

Standard Deviation of Transformed Data:

$$\begin{aligned} & \sqrt{\frac{((x_1 + c) - (\bar{x} + c))^2 + ((x_2 + c) - (\bar{x} + c))^2 + ((x_3 + c) - (\bar{x} + c))^2 + ((x_4 + c) - (\bar{x} + c))^2 + ((x_5 + c) - (\bar{x} + c))^2}{4}} \\ &= \sqrt{\frac{(x_1 + c - \bar{x} - c)^2 + (x_2 + c - \bar{x} - c)^2 + (x_3 + c - \bar{x} - c)^2 + (x_4 + c - \bar{x} - c)^2 + (x_5 + c - \bar{x} - c)^2}{4}} \\ &= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2}{4}} \\ &= s \end{aligned}$$

## 2.4 The Normal Distribution

---

### Objectives

- to learn to convert values to  $z$ -scores (standardize)
- to learn to convert  $z$ -scores to values in the original units (unstandardize)
- to find areas under the standard normal curve
- to use a table of the normal distribution to estimate proportions and probabilities of events that come from a population that is normally distributed
- to find the value that is located at a given percentile of the normal distribution

### Important Terms and Concepts

- standard normal curve
- $z$ -scores or standard units

### Lesson Planning

#### Class Time

Two to three days

#### Materials

None

### Suggested Assignments

Classwork		
Essential	Recommended	Optional
D39, D40, D42, D43, D45, D46 P38–P42, P44, P46–P48	D38, D41, D44 P37, P43, P45	

Homework		
Essential	Recommended	Optional
E46–E48, E52, E54, E56	E45, E49–E51, E53, E55	

## Lesson Notes: Solving the Unknown Problems and Calculator Use

With a graphing calculator, a student can circumvent most of the procedures in this section. How much of that process should still be taught in the introductory statistics course is an open question.

Consequently, your biggest decision in this lesson is deciding when to show students how to use their calculator to find areas under normal curves. If your students aren't taking the AP Statistics exam, once they understand the concept of a z-score as the number of standard deviations from the mean, you may want to move to the calculator right away. Specifically, once you finish P43, skip over P44–P45 and use the calculator from that point on (rather than z-scores and the table) to answer all questions.

Students who will be taking the AP Statistics exam should be able to work through the process step-by-step, including the use of the table, because they may be asked for the results of any one step of this process. However, it is even more important that they are able to use their calculator because it will be quicker and more accurate for most questions.

To find the area between the values  $x_1$  and  $x_2$  under a normal curve with mean  $\mu$  and standard deviation  $\sigma$  on a TI-83, use the command

$$\text{normalcdf}(x_1, x_2, \mu, \sigma)$$

To find the value  $x$  in a normal distribution with mean  $\mu$  and standard deviation  $\sigma$  that gives an area of  $P$  below  $x$ , use the command

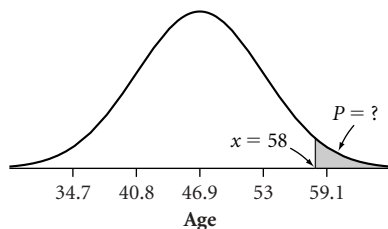
$$\text{invNorm}(P, \mu, \sigma)$$

If you omit values for  $\mu$  and  $\sigma$ , the calculator assumes they are 0 and 1.

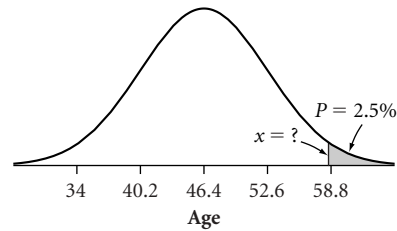
Answers in this section may differ slightly depending on whether students use a calculator or the table of the normal distribution.

### Discussion

**D38. a.** This is an unknown percentage problem.

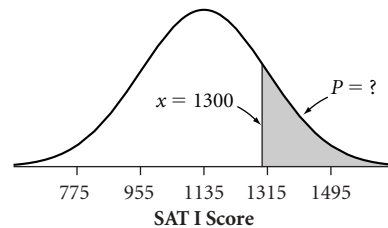


**b.** This is an unknown value problem. You need to find the age that cuts off the largest 2.5%.

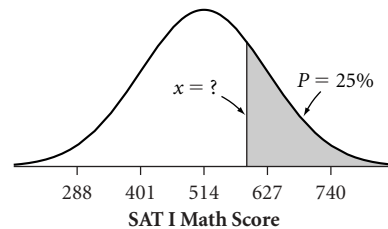


### Practice

**P37. a.** This is an unknown percentage problem. You need to find the percentage of scores over 1300.



**b.** This is an unknown value problem—you must find the value, not the percentage. You need to find the value that cuts off the bottom 75% of the distribution.

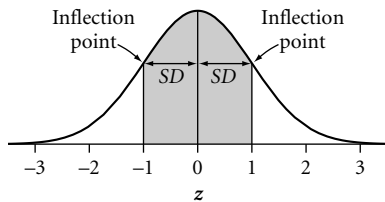


## Lesson Notes: The Standard Normal Distribution

All normal distributions can be transformed to a normal distribution with mean 0 and standard deviation 1 by the  $z$  transformation,

$$z = \frac{x - \text{mean}}{SD}$$

That is, when you compute a  $z$ -score, you are rescaling so that the mean of the new distribution is 0 and the standard deviation is 1. A  $z$ -score, then, tells how many standard deviations from the mean a value lies. This distribution, pictured on the next page, is called the *standard normal distribution*.



Students should make a sketch of the standard normal curve before they do any calculations.

Table A on page 701 of the appendix of the student text gives the areas in this standard normal distribution. After converting to  $z$ -scores, students can use this table to find the proportion of values that fall into a given region of any normal distribution.

### Discussion

- D39.** a. .8413 or 84.13%; .9943 or 99.43%  
 b. .1587 or 15.87%; .0057 or 0.57%  
 c. .9332 or 93.32%  
 d. .6827 or 68.27% (.6826 using Table A)

- D40.** a. 0  
 b. The lower quartile is the  $z$ -score that has 25% of the values below it. Looking in the center of the table, the  $z$ -score with a percentage closest to 25 is  $-0.67$ .  
 c. A percentage of 95 lies right between  $z$ -scores of 1.64 and 1.65, so the best answer is  $z = 1.645$ .  
 d.  $IQR = 1.349$  or 1.34

### Practice

- P38.** a.  $-0.47$       b.  $-0.23$   
 c. 1.13            d. 1.555
- P39.** a. .0129          b. .0475  
 c. .3446          d. .7881
- P40.** a.  $.9279 - .0721 = .8558$  or 85.58%.  
 b.  $.9987 - .0013 = .9974$  or 99.74%.
- P41.** a. The  $z$ -score that has 5% of the values below it is  $-1.645$ , and the  $z$ -score that has 5% of the values above it is 1.645. So the interval is  $-1.645$  to 1.645.  
 b.  $-1.96$  to 1.96

### Lesson Notes: Standard Units

The use of  $z$ -scores, or standard units, for a comparison is appropriate whenever you have two comparable normally distributed variables with different means and/or standard deviations.

Unless the distributions are approximately normal,  $z$ -scores should not be used to compare two values in the distribution. Percentiles, however, can always be used. (See E51 on page 70 in the student book.)

### Discussion

- D41.** a.  $\frac{(200 - 80)}{60} = 2$  hours  
 b. Subtract (recenter), and then divide (rescale). That is, how far from the exit? How many hours is that?

**D42.**

$$z_{heart} = \frac{90 - 289}{54} = -3.69$$

$$z_{cancer} = \frac{84 - 200}{31} = -3.74$$

The death rate for heart disease is 3.69 standard deviations below the mean. The death rate for cancer is 3.74 standard deviations below the mean. Thus, these rates are about equally extreme, but the death rate for cancer is slightly more extreme. (They are quite extreme because of Alaska's relatively young population.)

### Practice

- P42.** The death rate for cancer is more standard deviations below the mean, so it is a bit more extreme.

$$z_{heart} = \frac{240 - 289}{54} = -0.91$$

$$z_{cancer} = \frac{166 - 200}{31} = -1.10$$

- P43.** a. The death rate for cancer is more standard deviations above the mean, so it is more extreme.

$$z_{heart} = \frac{365 - 289}{54} = 1.41$$

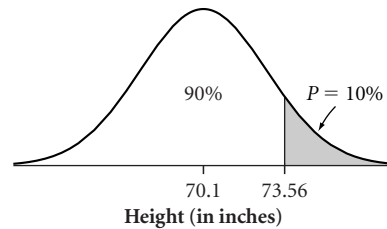
$$z_{cancer} = \frac{257 - 200}{31} = 1.84$$

- b. The death rate for heart disease in Colorado is more extreme than the death rate for cancer in Texas.

$$z_{heart} = \frac{184 - 289}{54} = -1.94$$

$$z_{cancer} = \frac{161 - 200}{31} = -1.26$$

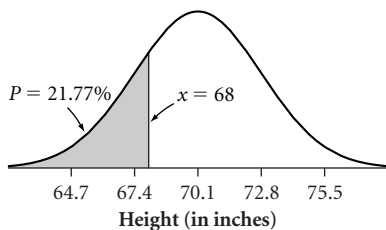
- P44. a. 2            b. 1            c. 1.5  
 d. 3            e. -1          f. -2.5
- P45. a. 30            b. 22  
 c. 85            d. -9.5



## Lesson Notes: Solving the Unknown Percentage Problem and the Unknown Value Problem

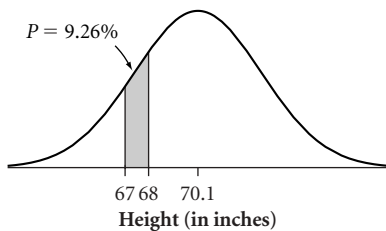
### Discussion

- D43. a. The z-score for the height 68 is  $\frac{(68 - 70.1)}{2.7} = -0.78$ . The area under the normal curve to the left of this point is .2177. Thus, about 21.77% of U.S. males between 18 and 24 are less than 68 inches tall.



- b. From part a, .2177 of the men are below the height of 68 inches. Similarly, the z-score for a height of 67 inches is -1.15, and so .1251 of the men are below that height. The proportion in between is  $.2177 - .1251 = .0926$ .

So about  $.0926(13,000,000) = 1,203,800$  are between the two heights.



- c. A percentile of 90 corresponds to a z-score of about 1.28. Using the formula,
- $$x = \text{mean} + z \cdot SD = 70.1 + 1.28(2.7) = 73.56$$

or, alternatively, solving

$$1.28 = \frac{x - 70.1}{2.7}$$

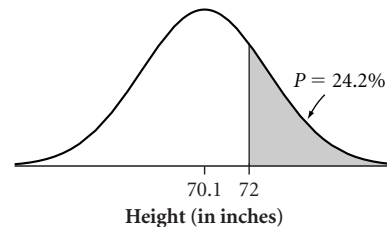
$$x = 73.56$$

- D44. The shape will not change. The mean will be  $\frac{70.1}{12} \approx 5.84$  feet, and the SD will be  $\frac{2.7}{12} = 0.225$  feet.

- D45.  $176 \pm (1.645)30$  or 126.65 mg/dl and 225.35 mg/dl.

### Practice

- P46. a. The z-score for the height 72 is  $\frac{(72 - 70.1)}{2.7} = 0.70$ . The area under the normal curve to the right of this point is  $1 - .7580 = .2420$ . Thus, about 24% of U.S. males between 18 and 24 are taller than 72 inches.



- b. The z-score for the 35th percentile is -0.385. The height that corresponds to that z-score is

$$x = \text{mean} + z \cdot SD = 64.8 + (-0.385)2.5 \approx 63.84 \text{ inches}$$

- P47.  $1100 \pm (1.96)180$  or roughly 747 and 1453

## Lesson Notes: Central Intervals for Normal Distributions

### Discussion

Some textbooks call this the “Empirical Rule.”

- D46. a.  $289 \pm 1.645(54)$  or about 200 to 378  
 b.  $289 \pm 1.96(54)$  or about 183 to 395  
 c.  $200 \pm 1.645(31)$  or about 149 to 251

### Practice

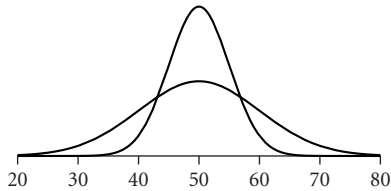
- P48. a. The z-score is -0.91, which is not outside either interval.  
 b. The z-score is -1.10, which is not outside either interval.

- c. The z-score is  $-3.69$ , which is outside both intervals.  
 d. The z-score is  $-3.74$ , which is outside both intervals.

### Exercises

E51 and E52 are important exercises to discuss as many students forget that z-scores can be used to estimate proportions only when the distribution is approximately normal.

E45.



E46. 68%; 95%; 16%; 84%; 97.5%; 2.5%

- E47. a. i. .6340 (calculator: .6319)  
 ii. .0392 (calculator: .0395)  
 iii. .3085 (calculator: .3101)

b. The middle 95% of scores range from, approximately,

$$505 - 1.96(111) \approx 287$$

to

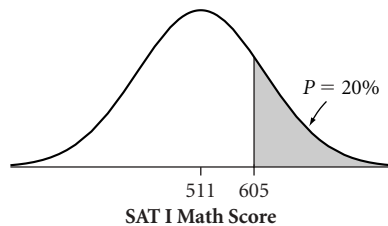
$$505 + 1.96(111) \approx 723$$

E48. The z-score that has an area of .80 below it is about  $z = .84$ .

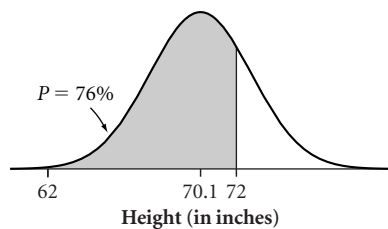
Unstandardizing,

$$x = \text{mean} + z \cdot \text{SD} = 511 + .84(112) \approx 605$$

The college should send letters to students who get 605 or more on the exam.



E49. First, find the percentage of men who qualify.

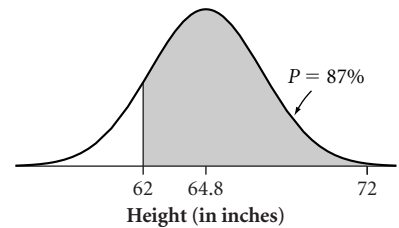


$$z = \frac{x - \text{mean}}{\text{SD}} = \frac{62 - 70.1}{2.7} = -3$$

$$z = \frac{x - \text{mean}}{\text{SD}} = \frac{72 - 70.1}{2.7} \approx 0.70$$

The area between these z-scores is about .76. About 76% of men aged 18 to 24 in the United States meet the height qualifications to be a flight attendant for United Airlines.

Next, find the percentage of women who qualify.



$$z = \frac{x - \text{mean}}{\text{SD}} = \frac{62 - 64.8}{2.5} = -1.12$$

$$z = \frac{x - \text{mean}}{\text{SD}} = \frac{72 - 64.8}{2.5} = 2.88$$

The area between these two z-scores is about .87.

About 87% of women aged 18 to 24 in the United States meet the height qualifications to be a flight attendant for United Airlines, a higher percentage than men.

- E50. a. About .0000270786, or 0.00270786%, are as tall or taller than Karl Malone. There are only about 352 men aged 18 to 24 who are at least as tall.  
 b. About .001717, or 0.1717%, are as tall or taller than Michael Jordan. There are only about 22,324 men aged 18 to 24 who are at least as tall.  
 c. About .00000017136, or 0.0000017136%, are as tall or taller than Shaquille O'Neal. You would expect to find less than 1 man (or .22) this tall in the 18- to 24-year-old age group.  
 d. The estimates will be too small.

E51. You cannot use the normal distribution to solve this problem because the distribution of ages of cars is not approximately normal. In fact, it is strongly skewed right.

E52. a. The distribution is probably skewed right because it's not possible for the length of a reign to be much more than 1 standard deviation below the mean.

b. The  $z$ -score for 0 is  $\frac{(0 - 18.5)}{15.4} = -1.20$ , so about .1151 of the reigns.

c. If all values in the distribution must be positive and two standard deviations or less below the mean is less than 0, the distribution isn't approximately normal.

E53. a. about 145 points

b. about 25 points

c. From the graph, the middle 95% of the values appear to lie between about 90 and about 200. Using the mean and standard deviation from parts a and b, this interval is about 95 to 195.

d. The  $z$ -score for 150 is 0.20, and the area to the right of this point is .4207. The  $z$ -score for 190 is 1.80, and the area to the right of this point is .0359. A weakness here is that next year's teams may not look like a random sample from the set of teams since 1939. Modern teams place more emphasis on scoring than did the teams from an earlier era.

E54. a. .1587

b. 8.16

c. Solving  $-1.34 = \frac{6 - \text{mean}}{3}$ , you get  $\text{mean} = 10.02$ .

d. Solving  $0.25 = \frac{12 - \text{SD}}{10}$ , you get  $SD = 8$ .

E55. a. The  $z$ -scores for the quartiles are  $\pm 0.67$ . Thus,  $Q_1 = 6.65$  and  $Q_3 = 13.35$ .

b. The mean must be 150 because it lies midway between the quartiles in a normal distribution. Then  $SD \approx 44.78$ .

c. Solving  $-0.67 = \frac{100 - \text{mean}}{10}$ , you get  $\text{mean} = 106.7$ . Then, because the quartiles are symmetric about the mean,  $Q_3 = 113.4$ .

d. Because the quartiles are symmetric about the mean,  $Q_1 = 9$ . Then  $SD \approx 1.5$ .

E56. a. Using the plot, a score of 425 appears to be at about the 22nd percentile. Assuming a normal distribution, the  $z$ -score for 425 is  $-0.72$ , giving a percentile of 23.58. These are fairly close.

b. From Display 2.55, the 40th percentile appears to be about 480. The 40th percentile under a standard normal curve has a  $z$ -score of  $-0.25$ , which translates to a test score of  $x = \text{mean} + z \cdot SD \approx 505 - 0.25(111) \approx 477$ . Again, these are quite close, which gives some evidence in support of the normality of the scores.

c. Answers will vary somewhat. From the plot, the median appears to be about 510. In a normal distribution, the median should be close to the mean, or 505. The mean and median here are close.

d. The quartiles are about 430 and about 585, giving an  $IQR$  of about  $585 - 430 = 155$ .

From the standard normal curve, the quartiles have  $z$ -scores of approximately  $-0.67$  and  $+0.67$  and the median has a  $z$ -score of 0. Thus, the approximate quartiles for the exam scores with mean 505 and standard deviation 111 are

$$Q_1 = 505 - 0.67(111) \approx 431$$

$$Q_3 = 505 + 0.67(111) \approx 579$$

The normal model is looking good!

## Review

Homework	
Essential	E57–E59, E62, E63, E65, E67, E69, E71
Recommended	E60, E66, E74
Optional	E61, E64, E68, E70, E72, E73, E75, E76

You may wish to save some or all of these exercises to use as review before the AP Statistics exam.

### Review Exercises

E57. a. Stem-and-leaf of Number     $N = 51$   
 Leaf Unit = 1.0

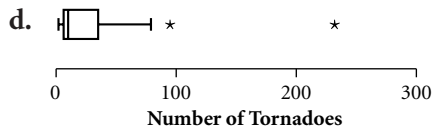
```

16  0  000000011122244
25  0  566778899
(5) 1  00013
21  1  5
20  2  244
17  2  66
15  3  03
13  3  579
10  4  2
9   4  58
7   5  14
5   5
5   6
5   6
5   7  2
4   7  69
HI  95, 232
0|5 stands for 5 tornadoes

```

- b. Minimum: 0  
 $Q_1$ : 2  
 Median: 10  
 $Q_3$ : 35  
 Maximum: 232

c. Outliers fall below  $2 - 1.5(35 - 2) = -47.5$  or above  $35 + 1.5(35 - 2) = 84.5$ . There are two outliers, Florida and Texas.



e. Both plots show the strong skewness in the data and the outlier of Texas. In the stemplot, you can see that half the states have less than 10 tornadoes. The cleanness of the boxplot makes it clear how much of an outlier Texas actually is. (Of course, it is a very large state, which helps explain why it is an outlier.) However, you can't see from the boxplot at all that so many states have at most 1 tornado. Because the stemplot has a reasonable number of values in it and is consequently easy to read while carrying almost the complete values, it is reasonable to select it as the most informative.

f. The distribution of numbers of tornadoes is strongly skewed right with two outliers, Florida at 95 and Texas at 232. The median number of tornadoes is 10 with the middle half of the states having between 2 and 35.

E58. Outliers would lie below  $1170 - 1.5(1340 - 1170) = 915$  or above  $1340 + 1.5(1340 - 1170) = 1595$ .

- E59. a. Median: 23 or 24 cents  
 b.  $Q_1$ : about 10 cents  
 $Q_3$ : about 70 cents  
 $IQR$ :  $70 - 10 = 60$  cents

c. Skewed right

d. No, because the data are obviously skewed, as you can see from the summary computed in parts a and b.

- E60. I. B      II. D  
 III. A      IV. C

E61. Important warning in this exercise: Percentiles are not *score/total* but a measure of position relative to the number of scores.

a. The mean of the scores on Test I is 15.5, and the standard deviation is 3.028. The

score of 19 has a z-score of 1.16 and is at the 90th percentile, relative to the other 9 scores.

b. The mean of the scores on Test II is 6.4, and the standard deviation is 8.708. The score of 18 has a z-score of 1.33 but is only at the 80th percentile, relative to the other 9 scores.

c. Answers will vary, and rightfully so. The student who got a 19 on Test I did better than all but one other student in the class. However, the student who got an 18 on Test II did *much* better than all but two students in the class.

E62. a. The state with the lowest average income in dollars in 1980 had an average income of \$6,926.

b. There are no outliers for 1980. There is at least one outlier for 1994 because  $Q_3 + 1.5(IQR) = 22,542 + 5,598 = \$28,140$ , which is less than the maximum of \$30,721. Thus, the state with the maximum income is an outlier. There may be other states as well on the high end that are outliers.

c. No. Alabama remains below the lower quartile of the distribution, but you cannot say exactly where in the lowest quarter it lies. It is tempting to find Alabama's relative position using z-scores, but there is no indication that these incomes are normally distributed and that it is appropriate to use z-scores for comparison only in that case.

E63. The stems represent the first three digits of the year. The leaves represent the final digit of the year. The distribution of record lows is relatively uniform over the years since 1890. The distribution of record highs is also relatively uniform, except for a spike during the 1930s, when records were set for low temperatures in many states. Those years were also overrepresented in record highs. The center and spread of the two distributions are about the same.

E64. a. Each value is the instructional time in mathematics in a given school. (Only schools that have all 9th graders in the same mathematics course are plotted.)

b. Norway and Singapore have no variation by school in instructional time in mathematics. These countries may mandate instructional time for every school in the country.

c. The median instructional time in mathematics is about the same for the United States as the overall median. In addition, the variation from school to school is about average for the countries in the plot. There are no schools in the United States with an unusually large or small amount of instructional time.

**E65. a.** Guesses may vary. In actual fact, Region 1 is Africa, Region 2 is South and Central America, Region 3 is Asia, and Region 4 is Europe. The outlier for Asia is Bangladesh. The outlier for Africa is Angola (Egypt is the country with 99% in Africa).

**b.** Distributions 1, 2, and 4 are skewed left.

**c.** The dot plots are in the same order as the boxplots, A being Africa, B South and Central America, C Asia, and D Europe.

**d.** The outlier shown in the boxplot in Region 1 doesn't appear to be an outlier in the dot plot. Region 4 does not look skewed in the dot plot even though it appears so in the boxplot. The number of countries plotted is small, and the values vary a lot, so the locations of the quartiles might change quite a bit with small changes in the data. Dot plots give the better picture here.

**E66.** The mean grade is 2.98 with a standard deviation of 1.33.

**E67. a.** The median number of deaths is 47. Half of the cities have fewer than 47 pedestrian deaths per year, and half have more.

**b.** The quartiles are 28.5 and 88.5. An outlier is a city whose number of deaths exceeds  $88.5 + 1.5(88.5 - 28.5) = 178.5$  or is less than  $28.5 - 1.5(88.5 - 28.5) = -61.5$ , which is impossible. There are three outliers, Chicago, Los Angeles, and New York. One explanation is the large populations these cities have; these are the three largest cities in the United States.

**c.** Student plots and explanations will vary. A stemplot is a very good choice. It reveals the three outliers and shows that the distribution is skewed right.

Stem-and-leaf of NumDeath N = 41

Leaf Unit = 1.0

```

2   1   79
11  2  023457889
19  3  33456777
(3) 4   378
19  5   118
16  6    6
15  7   69
13  8   045
10  9   268
7   10  017
4   11
4   12  0

```

HI 180, 299, 310

11|7 represents 17 deaths

**d.** Note what happened in part b. If the data were presented in rates per 100,000 population, these three cities might not be outliers. In fact, their rates of pedestrian death might be relatively small.

**E68. a.** 1267

**b.** Bacon himself

**c.** For Bacon and for Connery,  $n = 263,484$ . The mean Bacon number is 2.8. The mean Connery number is only 2.6. Connery is the better center. Also, Connery has only 974 actors who are outliers, that is, those with Connery numbers of 5–7 (not counting the outlier “0,” himself), whereas Bacon has 2237 outliers (those with Bacon numbers of 5–8) not counting himself. That is, there are more people at a far remove from Bacon than at a far remove from Connery.

**d.** Students should realize that the numbers must be equal. In fact, it is 2: Sean Connery was in *The Untouchables* (1987) with Kevin Costner, who was in *JFK* (1991) with Kevin Bacon.

**E69. a.** Except for a slight bulge around .220, the batting averages look quite normal in their distribution.

**b.** The mean is about .270, and the standard deviation is about .030.

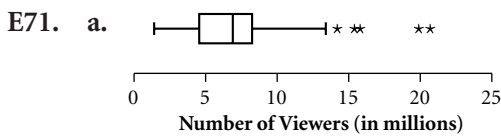
**c.** about .200 to .320

**E70. a.** Again, the histogram of batting averages looks quite normal in shape with center at about .260. The standard deviation is approximately .040. Quite a few regular players hit under .200 in the National League. In fact, .200 has a  $z$ -score of  $-1.5$ , and this is about the 7th percentile of a normal density curve.

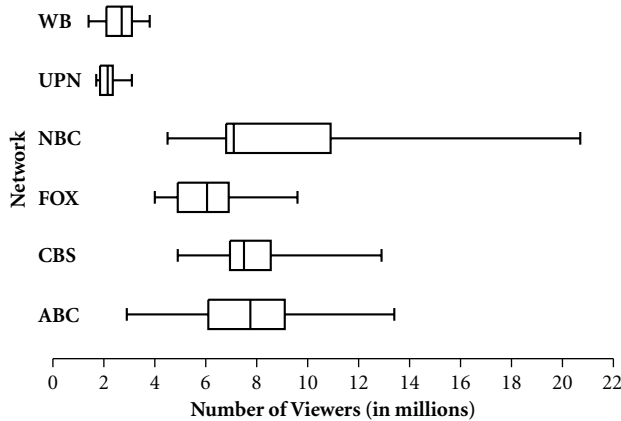
**b.** The batting averages in both leagues have distributions that are approximately normal in shape. The American League has a higher mean (by about .010) and less spread.

**c.** As seen above, the  $z$ -score corresponding to .200 in the National League is  $z = -1.5$ . The corresponding batting average,  $x$ , in the American League would still be 1.5 standard deviations below the mean, or

$$x = .270 - (1.5)(.030) = .225$$



**b.** These boxplots are made with Fathom.



**E72.** No, this is true only for normal distributions. For example, the set of values  $\{2, 2, 4, 6, 8, 8\}$  is symmetric and has mean and median both equal to 5. The standard deviation is about 2.76. Only two, or one-third, of the values are within one standard deviation of the mean.

**E73.** There are many possible responses. An example is  $\{1, 1, 1, 1, 1, 2, 2, 10\}$ , which has mean 2.375 and standard deviation of about 3.11. One standard deviation below the mean is less than 0.

**E74. a.** Developing countries have lower life expectancies than do developed countries. Thus, Region 1 must be Africa (developing countries, for the most part) and Region 3 must be Europe (developed countries). The Middle East, Region 2, has a mixture of developed and developing countries.

**b.** A is for Region 3 (Europe); B is for Region 1 (Africa); C is for Region 2 (Middle East).

**E75. a.** With *Seinfeld*, the midrange is  $\frac{(2.32 + 76.26)}{2} = 39.29$ . Without *Seinfeld*, the midrange is  $\frac{(2.32 + 58.53)}{2} = 30.425$ . The midrange is not resistant and is extremely sensitive to outliers because it is computed using only the maximum and minimum.

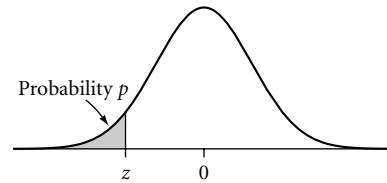
**b.** The total of the ratings without the *Seinfeld* episode is  $(101)11.187 - 76.26 = 1053.627$ . So the mean rating is  $\frac{1053.627}{100} \approx 10.54$ .

**E76. a.** Students will need to remove the top 2 longevities and the bottom 2 longevities from the data set and compute the mean of the remaining 35 animals. The trimmed mean is 30.5114.

**b.** Yes, because most outliers are removed before computing the trimmed mean.

# Table A Standard Normal Probabilities

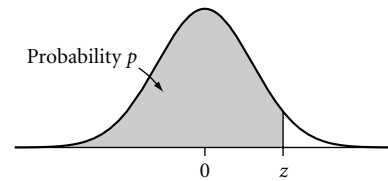
Table entry for  $z$  is the probability lying below  $z$ .



$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.8	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.7	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

# Table A Standard Normal Probabilities (continued)

Table entry for  $z$  is the probability lying below  $z$ .



$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999