Quantitative Literacy:
Thinking Between the Lines

Crauder, Noell, Evans, Johnson

Chapter 6: Statistics

Chapter 6: Statistics

Lesson Plan

- Data summary and presentation: Boiling down the numbers
- The normal distribution: Why the bell curve?
- The statistics of polling: Can we believe the polls?
- Statistical inference and clinical trials: Effective drugs?

Learning Objectives:

- Know the statistical terms used to summarize data
- Calculate mean, median, and mode
- Understand the five-number summary and boxplots
- Calculate the standard deviation
- Understand histograms

6.1 Data summary and presentation: Boiling down the numbers

- The **mean (average)** of a list of numbers is the sum of the numbers divided by the number of entries in the list.
- The **median** of a list of numbers is the middle number, the middle data point. If there is an even number of data points, take the average of the middle two numbers.
- The **mode** is the most frequently occurring data points. If there are two such numbers, the data set is called **bimodal.**
- If there are more than two such numbers, the data set is **multimodal.**
- If no number repeats, the data set has **no mode.**

6.1 Data summary and presentation: Boiling down the numbers

- **Example:** The Chelsea Football Club (FC) is a British soccer team. The following table shows the goals scored in the games played by Chelsea FC between September 2007 and May 2008. The data are arranged according to the total number of goals scored in each game.

Goals scored by either team	0	1	2	3	4	5	6	7	8
Number of games	7	14	20	11	3	2	1	2	2

- Find the mean, median, and mode for the number of goals scored per game.

- **Solution:**
  - To find the mean, we add the data values (the total number of goals scored) and divide by the number of data points.
  - To find the total number of goals scored, for each entry we multiply the goals scored by the corresponding number of games. Then we add.
  - The total number of goals scored:

$$(7 \times 0) + (14 \times 1) + (20 \times 2) + (11 \times 3) + (3 \times 4) + (2 \times 5) + (1 \times 6) + (2 \times 7) + (2 \times 8) = 145$$

  - The number of data points or the total number of games:

$$7 + 14 + 20 + 11 + 3 + 2 + 1 + 2 + 2 = 62$$

- **Solution (cont.):**
  - The mean: the total number of goals scored divided by the number of games played:

$$\text{Mean} = \frac{145}{62} = 2.3$$

  - The median: the total number of games is 62, which is even, so we count from the bottom to find the 31st and 32nd lowest total goal scores. These are both 2:

$$\text{Median} = 2$$

  - The mode: because 2 occurs most frequently as the number of goals (20 times): $\text{Mode} = 2$
  - Thus on average, the teams combined to score 2.3 goals per game. Half of the games had goals totaling 2 or more, and the most common number of goals scored in a Chelsea FC game was 2.

6.1 Data summary and presentation: Boiling down the numbers

- **Example:** The following list gives home prices (in thousands of dollars) in a small town:

$$80, 120, 125, 140, 180, 190, 820$$

The list includes the price of one luxury home. Calculate the mean and median of this data set. Which of the two is more appropriate for describing the housing market?

- **Solution:** $\text{Mean} = \frac{80+120+125+140+180+190+820}{7} = \frac{1655}{7}$

Or about 236.4 thousand dollars. This is the average price of a home.

The list of seven prices is arranged in order, so the median is the fourth value, 140 thousand dollars.

Note that the mean is higher than the cost of every home on the market except for one—the luxury home. The median of 140 thousand dollars is more representative of the market.

6.1 Data summary and presentation: Boiling down the numbers

- An **outlier** is a data point that is significantly different from most of the data.

- The **first quartile** of a list of numbers is the median of the lower half of the numbers in the list.
- The **second quartile** is the same as the median of the list.
- The **third quartile** is the median of the upper half of the numbers in the list.
- The **five-number summary** of a list of numbers consists of the minimum, the first quartile, the median, the third quartile, and the maximum.

6.1 Data summary and presentation: Boiling down the numbers

- **Example:** Each year *Forbes* magazine publishes a list it calls the Celebrity 100. The accompanying table shows the top nine names on the list for 2009, ordered according to the ranking of *Forbes*. The table also gives the incomes of the celebrities between June 2008 and June 2009.
- Calculate the five-number summary for this list of incomes.

6.1 Data summary and presentation: Boiling down the numbers

Celebrity	**Income** (millions of dollars)
Angelina Jolie	27
Oprah Winfrey	275
Madonna	110
Beyonce Knowles	87
Tiger Woods	110
Bruce Springsteen	70
Steven Spielberg	150
Jennifer Aniston	25
Brad Pitt	28

- **Solution:** First we arrange the incomes in order:

$$25 \;\; 27 \;\; 28 \;\; 70 \;\; 87 \;\; 110 \;\; 110 \;\; 150 \;\; 275$$

- The lower half of the list consists of the four numbers less than the median ($87 million), which are:

$$25 \;\; 27 \;\; 28 \;\; 70$$

  The median of this lower half is 27.5, so the first quartile of incomes is $27.5 million.

- The upper half of the list consists of the four numbers greater than the median, which are:

$$110 \;\; 110 \;\; 150 \;\; 275$$

  The median of this upper half is 130, so the third quartile of incomes is $130 million.

- **Solution (cont.):**

Thus, the five-number summary is:

$$\begin{aligned}\text{Minimum} &= \$25 \text{ million}\\ \text{First quartile} &= \$27.5 \text{ million}\\ \text{Median} &= \$87 \text{ million}\\ \text{Third quartile} &= \$130 \text{ million}\\ \text{Maximum} &= \$275 \text{ million}\end{aligned}$$

6.1 Data summary and presentation: Boiling down the numbers

- **Boxplots:** There is a commonly used pictorial display of the five-number summary known as a *boxplot* (also called a *box and whisker diagram*). Figure 6.1 shows the basic geometric figure used in a boxplot.

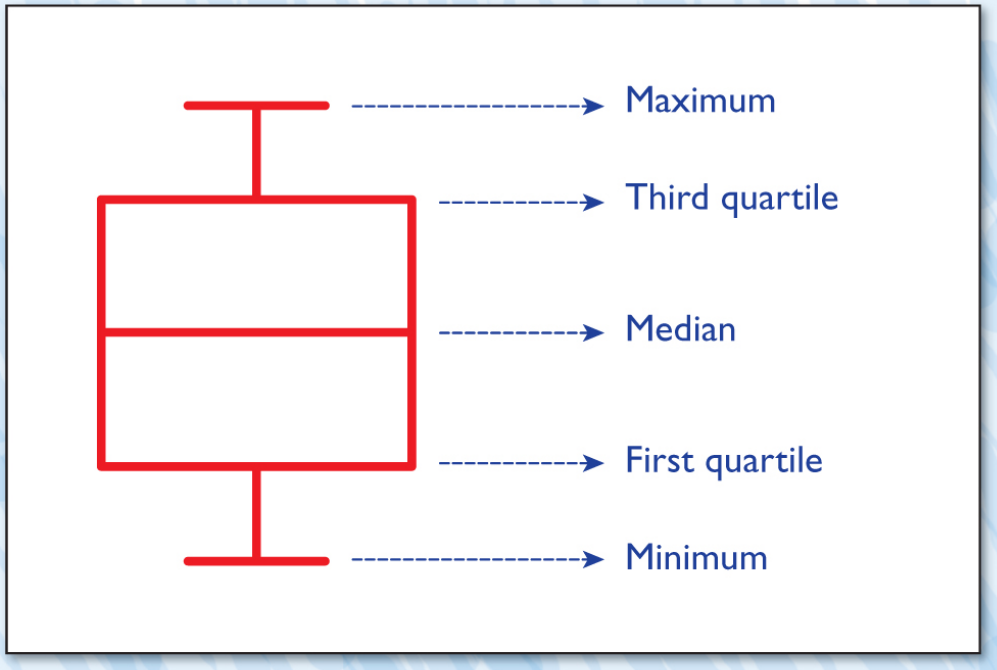

FIGURE 6.1 The basic boxplot diagram.

6.1 Data summary and presentation: Boiling down the numbers

- **Example:** A report on greenercars.org shows 2011 model cars with the best fuel economy.

1. Find the five-number summary for city mileage.
2. Present a boxplot of city mileage.
3. Comment on how the data are distributed about the median.

6.1 Data summary and presentation: Boiling down the numbers

Model	City Mileage (mpg)	Highway Mileage (mpg)
Toyota Prius	51	48
Honda Civic Hybrid	40	43
Honda CR-Z	35	39
Toyota Yaris	29	35
Audi A3	30	42
Hyundai Sonata	22	35
Hyundai Tucson	23	31
Chevrolet Equinox	22	32
Kia Rondo	20	27
Chevrolet Colorado/GMC Canyon	18	25

- **Solution:**

1. The list for city mileage, in order from lowest to highest:

$$18, 20, 22, 22, \mathbf{23}, \mathbf{29}, 30, 35, 40, 51$$

To find the median, we average the two numbers in the middle:

$$\text{Median} = \frac{23 + 29}{2} = 26 \text{ mpg}$$

The lower half of the list is $18, 20, 22, 22, 23$, and the median of this half is 22. Thus, the first quartile is 22 mpg.

The upper half of the list is $29, 30, 35, 40, 51$, and the median of this half is 35. Thus, the third quartile is 35 mpg.

- **Solution (cont.):**

2. The corresponding boxplot appears in Figure 6.2. The vertical axis is the mileage measured in miles per gallon.

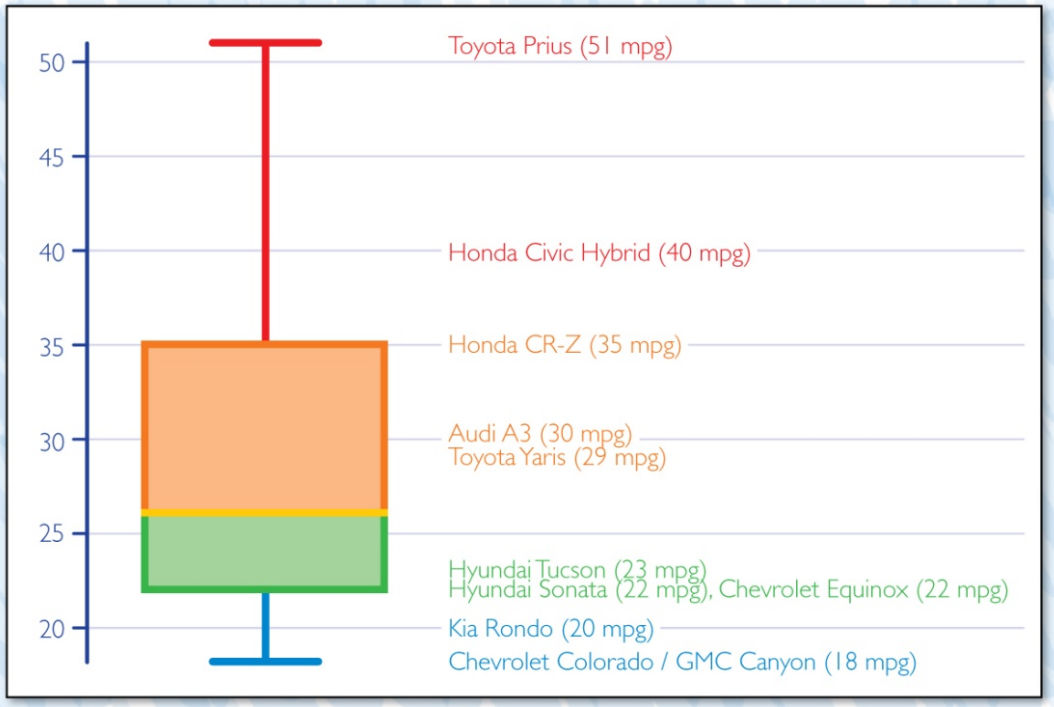

FIGURE 6.2 Boxplot for city mileage.

- **Solution (cont.):**

3. Referring to the boxplot, we note that the first quartile is not far above the minimum, and the median is barely above the first quartile. The third quartile is well above the median, and the maximum is well above the third quartile. This emphasizes the dramatic difference between the high-mileage cars (the hybrids) and ordinary cars.

6.1 Data summary and presentation: Boiling down the numbers

- The **standard deviation** is a measure of how much the data are spread out from the mean. The smaller the standard deviation, the more closely the data clustered about the mean.

Standard Deviation Formula

Suppose the data points are:

$$x_1, x_2, x_3, \ldots, x_n,$$

the formula for the standard deviation is

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}}$$

Where the Greek letter μ (mew) denotes the mean.

Calculating Standard Deviation

To find the standard deviation of n data points, we first calculate the mean μ. The next step is to complete the following calculation template:

Data	Deviation	Square of deviation
⋮	⋮	⋮
x_i	$x_i - \mu$	Square of second column
⋮	⋮	⋮
		Sum of third column Divide the above sum by n and take the square root.

6.1 Data summary and presentation: Boiling down the numbers

- **Example:** Two leading pitchers in Major League Baseball for 2011 were Roy Halladay of the Philadelphia Phillies and Felix Hernandez of the Seattle Mariners. Their ERA (Earned Run Average—the lower the number, the better) histories are given in the table below.

Pitcher	ERA 2006	ERA 2007	ERA 2008	ERA 2009	ERA 2010
R. Halladay	3.19	3.71	2.78	2.79	2.44
F. Hernandez	4.52	3.92	3.45	2.49	2.27

Calculate the mean and the standard deviation for Hallady's ERA history. It turns out that the mean and standard deviation for Hernandez's ERA history are $\mu = 3.33$ and $\sigma = 0.85$. What comparisons between Halladay and Hernandez can you make based on these numbers?

- **Solution (cont.):** We conclude that the mean and the standard deviation for Halladay's ERA history are $\mu = 2.98$ and $\sigma = 0.43$.
- Because Halladay's mean is smaller than Hernandez's mean of $\mu = 3.33$, over this period Halladay had a better pitching record.
- Halladay's ERA had a smaller standard deviation than that of Hernandez (who had $\sigma = 0.85$), so Halladay was more consistent—his numbers are not spread as far from the mean.

6.1 Data summary and presentation: Boiling down the numbers

- **Example:** Below is a table showing the Eastern Conference NBA team free-throw percentages at home and away for the 2007–2008 season. At the bottom of the table, we have displayed the mean and standard deviation for each data set.

  What do these values for the mean and standard deviation tell us about free-throws shot at home compared with free-throws shot away from home?

  Does comparison of the minimum and maximum of each of the data sets support your conclusions?

6.1 Data summary and presentation: Boiling down the numbers

Team	Free-throw percentage at home	Free-throw percentage away	Team	Free-throw percentage at home	Free-throw percentage away
Toronto	81.2	77.6	Milwaukee	73.3	76.6
Washington	78.2	75.4	Miami	72.7	75.5
Atlanta	77.2	75.2	New York	72.7	73.9
Boston	77.1	74.3	Orlando	72.1	75.4
Indiana	76.8	75.7	Cleveland	71.7	74.8
Detroit	76.7	74.4	Charlotte	71.4	74.7
Chicago	75.6	76.6	Philadelphia	70.6	77.2
New Jersey	73.6	76.8			
Mean	74.73	75.61			
Standard deviation	2.95	1.09			

- **Solution:** The means for free-throw percentages are 74.73 at home and 75.61 away, so on average the teams do somewhat better on the road than at home.
  - The standard deviation for home is 2.95 percentage points, which is considerably larger than the standard deviation of 1.09 percentage points away from home. This means that the free-throw percentages at home vary from the mean much more than the free-throw percentages away.
  - The difference between the maximum and minimum percentages shows the same thing: The free-throw percentages at home range from 70.6 to 81.2%, and the free-throw percentages away range from 73.9% to 77.6%.

6.1 Data summary and presentation: Boiling down the numbers

- **Solution (cont.):** The plots of the data in figures 6.3 and 6.4 provide a visual verification that the data for home free-throws are more broadly dispersed than the data for away free-throws.

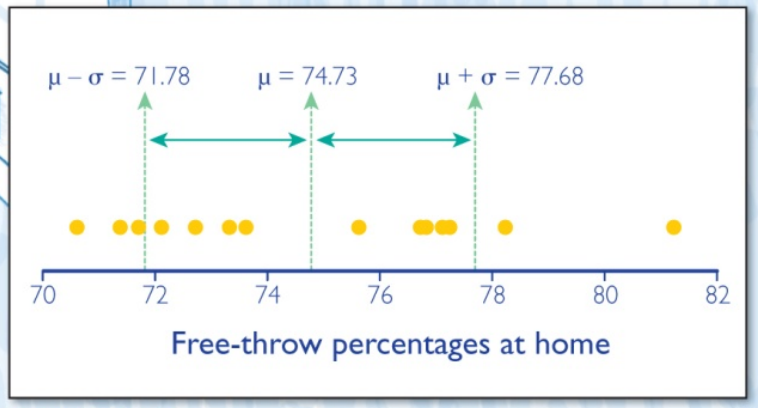

FIGURE 6.3 Eastern Conference NBA free-throw percentages at home.

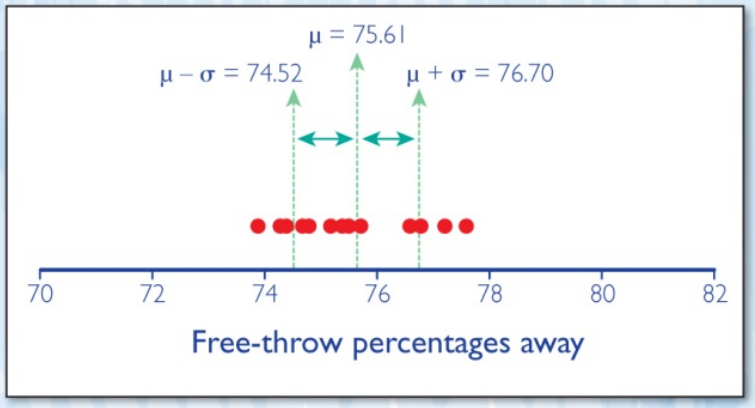

FIGURE 6.4 Eastern Conference NBA free-throw percentages away.

6.1 Data summary and presentation: Boiling down the numbers

- A **histogram** is a bar graph that shows the frequencies with which certain data occur.
- **Example:** Suppose we toss 1000 coins and write down the number of heads we got. We do this experiment a total of 1000 times. The accompanying table shows one part of the results from doing these experiments using a computer simulation.

Number of heads	451	457	458	459	461	462	463	464	465	467
Number of tosses (out of 1000)	2	2	1	3	3	2	1	3	1	1

The first entry shows that twice we got 451 heads, twice we got 457 heads, once we got 458 heads, and so on. The raw data are hard to digest because there are so many data points. The five-number summary provides one way to analyze the data.

An alternative way to get the data is to arrange them in groups and then draw a histogram.

6.1 Data summary and presentation: Boiling down the numbers

- **Example (cont.):** Suppose it turns out that the number of tosses yielding fewer than 470 heads is 23. Because 470 out of 1000 is 47%, it means that 23 tosses yields less than 47% heads. We find the accompanying table by dividing the data into groups this way.

Percent heads	Less than 47%	47% to 48%	48% to 49%	49% to 50%
Number of tosses	23	75	140	234
Percent heads	50% to 51%	51% to 52%	52% to 53%	At least 53%
Number of tosses	250	157	94	27

6.1 Data summary and presentation: Boiling down the numbers

- **Example (cont.):** Figure 6.7 shows a histogram for this grouping of the data. We can clearly see that the vast majority of the tosses were between 47% and 53% heads.

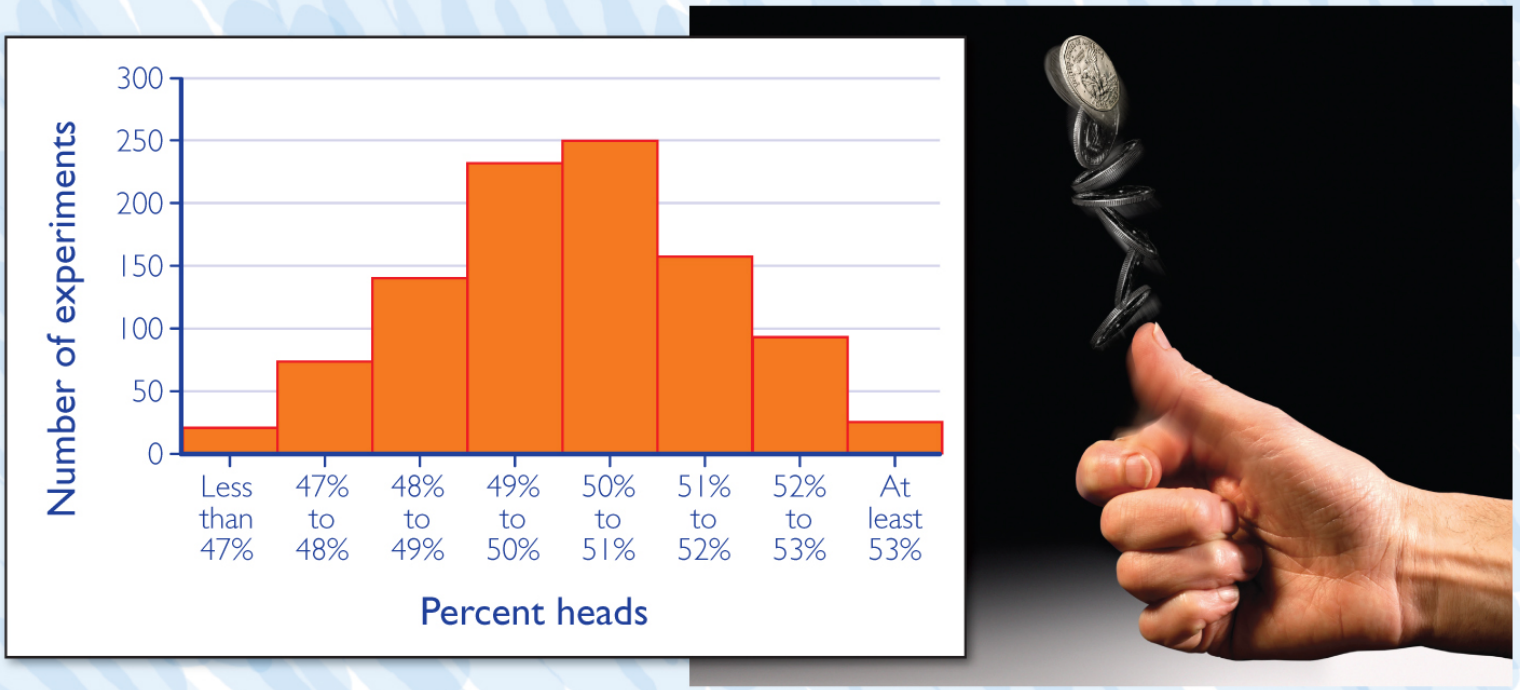

FIGURE 6.7 A histogram of coin tosses.

Learning Objectives:

- Understand why the normal distribution is so important.
  - The bell-shaped curve
  - Mean and standard deviation for the normal distribution
  - z-scores with Percentile scores
  - The Central Limit Theorem
  - Significance of apparently small deviations

6.2 The normal distribution: Why the bell curve?

- **The bell-shaped curve:** Figure 6.13 shows the distribution of heights of adult males in the United States. A graph shaped like this one resembles a bell—thus the *bell curve*. This bell-shaped graph is typical of normally distributed data.

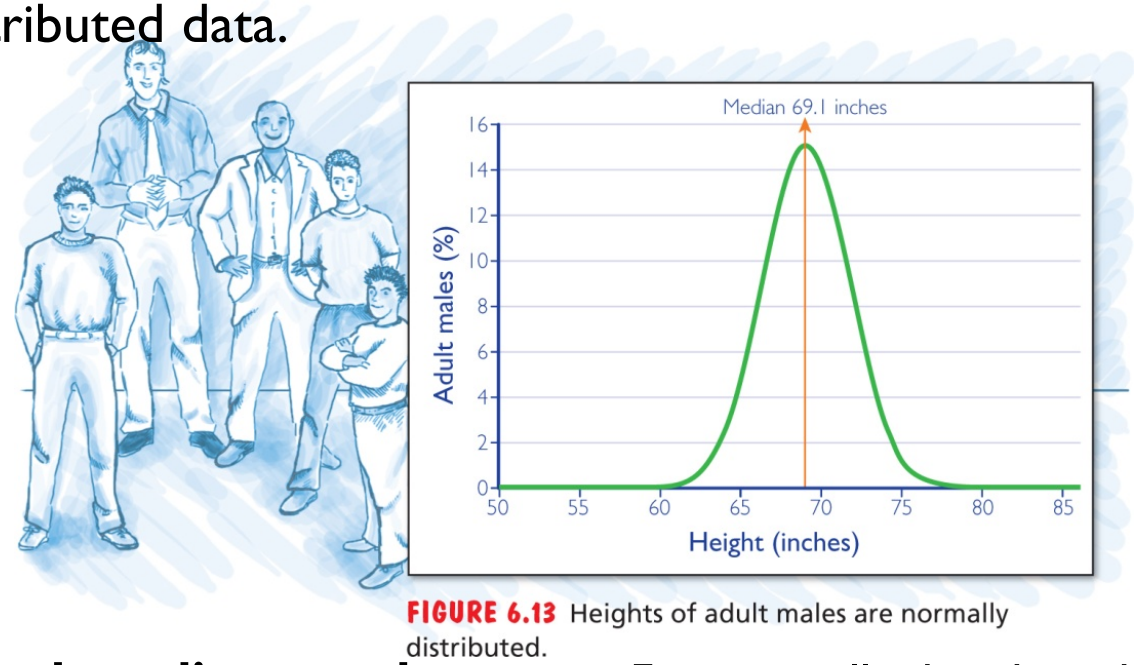

FIGURE 6.13 Heights of adult males are normally distributed.

- **The mean and median are the same:** For normally distributed data, the mean and median are the same. Figure 6.13 indicates that the median height of adult males is 69.1 inches. The average height of adult males is 69.1 inches.

6.2 The normal distribution: Why the bell curve?

- **Most data are clustered about the mean:** The vast majority of adult males are within a few inches of the mean.

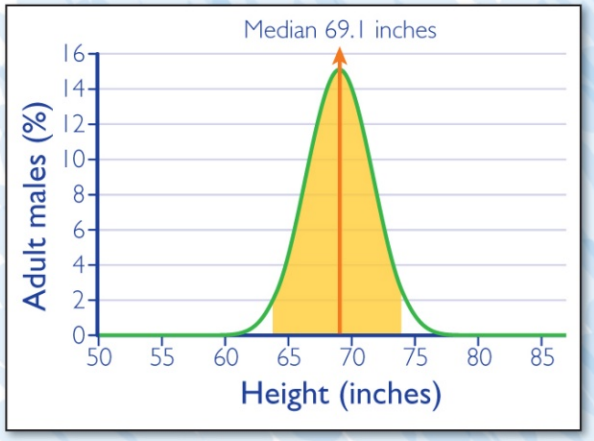

FIGURE 6.14 95% of adult males are within 5 inches of the median.

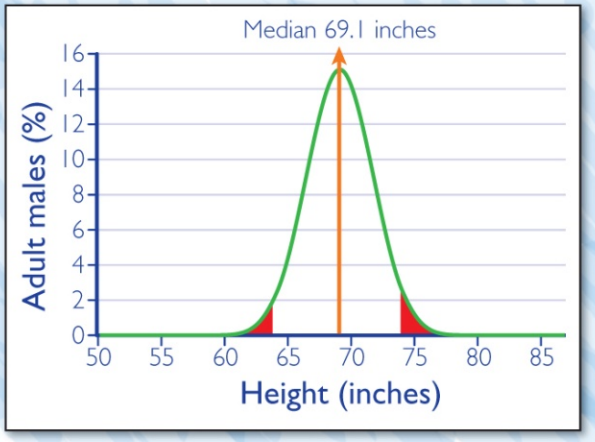

FIGURE 6.15 Relatively few men are very tall or very short.

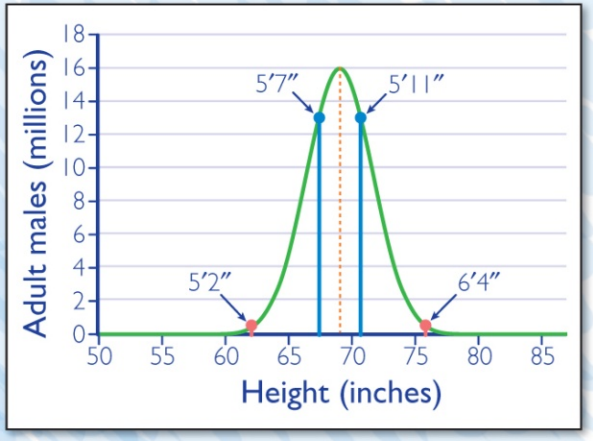

FIGURE 6.16 The bell curve is symmetric about the mean.

6.2 The normal distribution: Why the bell curve?

- **The bell curve is symmetric about the mean:** The curve to the left of the mean is a mirror image of the curve to the right of the mean. In terms of heights, there are about the same number of men 2 inches taller than the mean as there are men 2 inches shorter than the mean. This is illustrated in Figure 6.16.

- If data are **normally distributed:**

1. Their graph is a bell-shaped curve.
2. The mean and median are the same.
3. Most of the data tend to be clustered relatively near the mean.
4. The data are symmetrically distributed above and below the mean.

6.2 The normal distribution: Why the bell curve?

- **Example:** Figure 6.17 shows the distribution of IQ scores, and Figure 6.18 shows the percentage of American families and level of income. Which of these data sets appear to be normally distributed, and why?

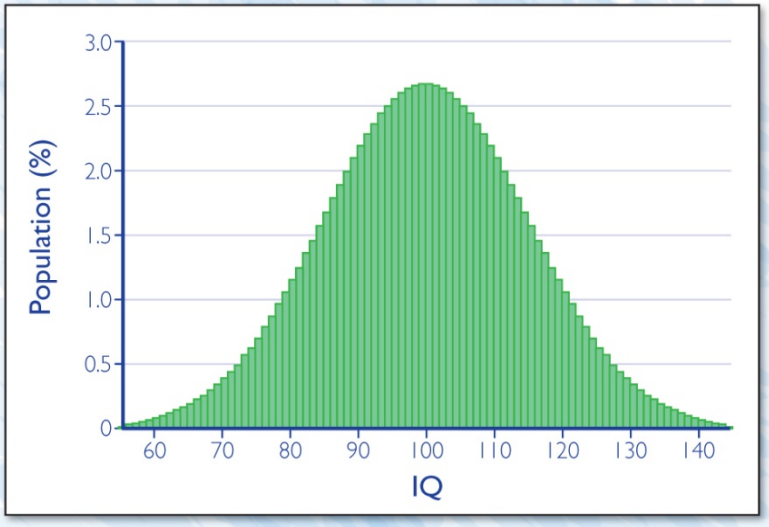

FIGURE 6.17 Percentage of population with given IQ score.

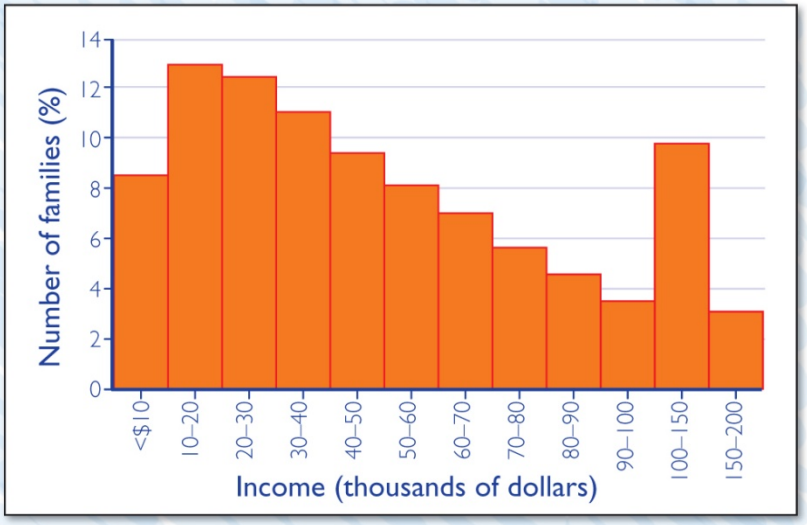

FIGURE 6.18 Percentage of families with given income.

- **Solution:** The IQ scores appear to be normally distributed because they are symmetric about the median score of 100, and most of the data relatively close to this value.
- Family incomes do not appear to be normally distributed because they are not symmetric. They are skewed toward the lower end of the scale, meaning there are many more families with low incomes than with high incomes.

6.2 The normal distribution: Why the bell curve?

- **Mean and standard deviation for the normal distribution:** A normal distribution, the mean and standard deviation completely determine the bell shape for the graph of the data.
- The mean determines the middle of the bell curve.
- The standard deviation determines how steep the curve is.
- A large standard deviation results in a very wide bell, and small standard deviation results in a thin, steep bell.

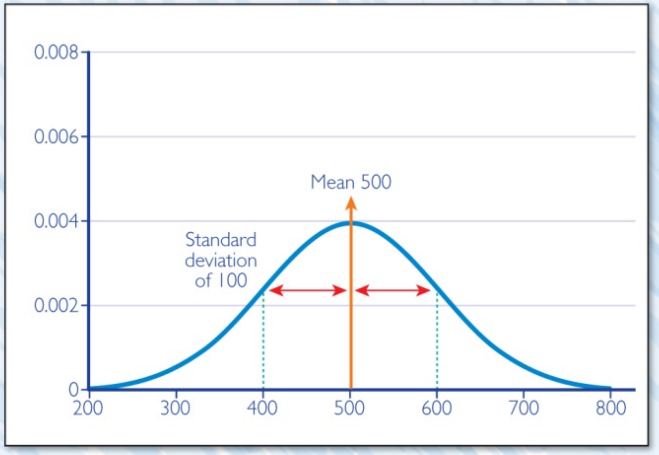

FIGURE 6.21 Normal curve with mean 500 and standard deviation 100.

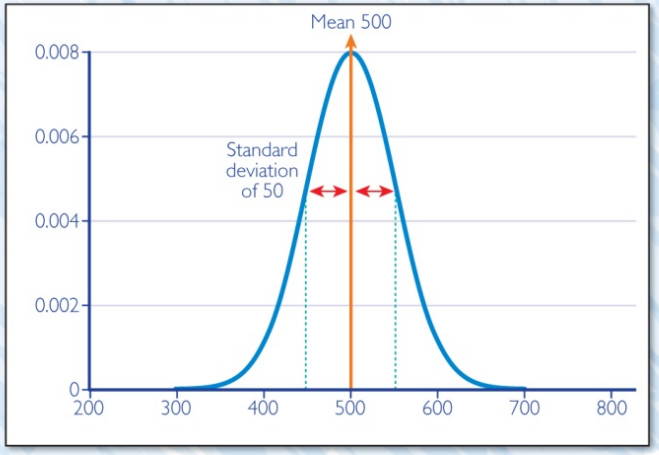

FIGURE 6.22 Normal curve with mean 500 and standard deviation 50.

6.2 The normal distribution: Why the bell curve?

Normal Data: 68-95-99.7% Rule

If a set of data is normally distributed:

- About 68% of the data lie within one standard deviation of the mean (34% within one standard deviation above the mean and 34% within one standard deviation below the mean). See Figure 6.23.
- About 95% of the data lie within two standard deviations of the mean (47.5% within two standard deviations above the mean and 47.5% within two standard deviations below the mean). See Figure 6.24.
- About 99.7% of the data lie within three standard deviations of the mean (49.85% within three standard deviations above the mean and 49.85% within three standard deviations below the mean). See Figure 6.25.

6.2 The normal distribution: Why the bell curve?

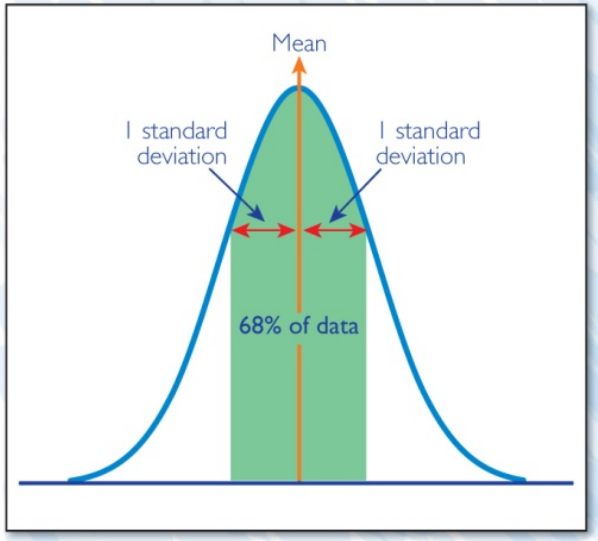

FIGURE 6.23 68% of data lie within one standard deviation of the mean.

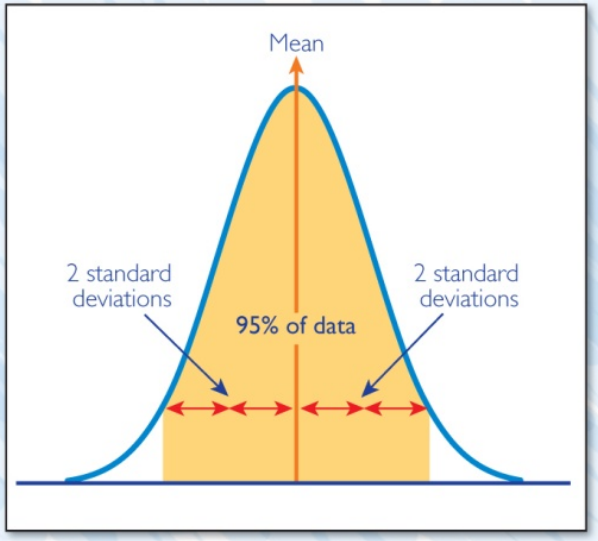

FIGURE 6.24 95% of data lie within two standard deviations of the mean.

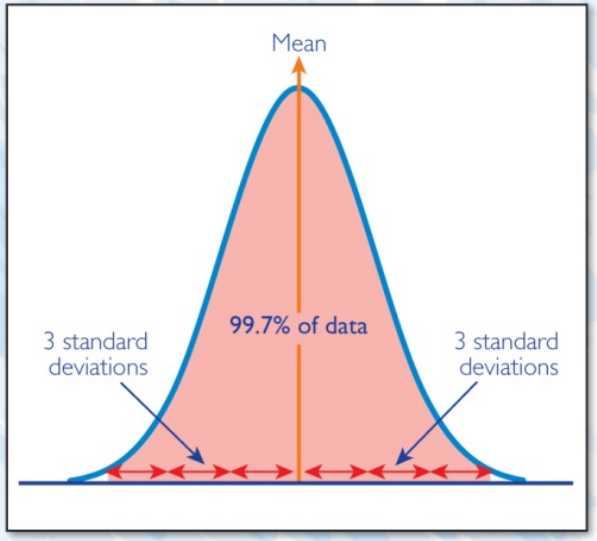

FIGURE 6.25 99.7% of the data lie within three standard deviations of the mean.

6.2 The normal distribution: Why the bell curve?

- **Example:** We noted earlier that adult male heights in the United States are normally distributed, with a mean of 69.1 inches. The standard deviation is 2.65 inches.

  What dose the 68-95-99.7% rule tell us about the heights of adult males?

- **Solution:** 68% of adult males are between

  $$69.1 - 2.65 = 66.45 \text{ inches (5 feet 6.45 inches) and}$$
  $$69.1 + 2.65 = 71.75 \text{ inches (5 feet 11.75 inches) tall}$$

- 95% are between

  $$69.1 - (2 \times 2.65) = 63.8 \text{ inches and}$$
  $$69.1 + (2 \times 2.65) = 74.4 \text{ inches tall}$$

- 99.7% are between

  $$69.1 - (3 \times 2.65) = 61.15 \text{ inches and}$$
  $$69.1 + (3 \times 2.65) = 77.05 \text{ inches tall}$$

6.2 The normal distribution: Why the bell curve?

- **Example:** The weights of apples in the fall harvest are normally distributed, with a mean weight of 200 grams and standard deviation of 12 grams. Figure 6.28 shows the weight distribution of 2000 apples. In a supply of 2000 apples, how many will weigh between 176 and 224 grams?

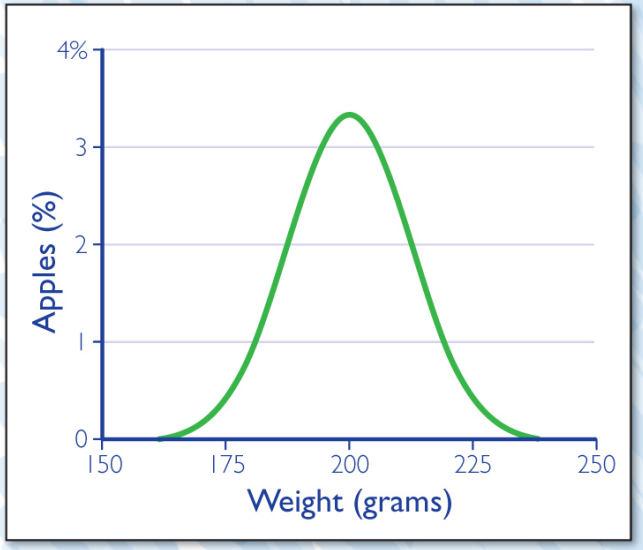

FIGURE 6.28 Apples with mean 200 grams and standard deviation 12 grams.

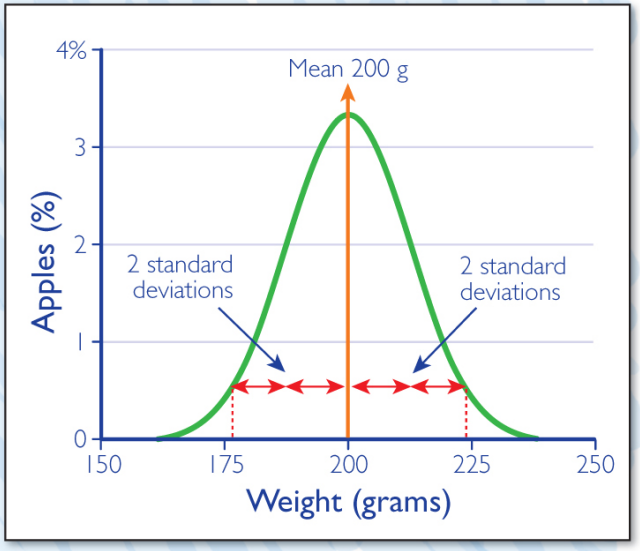

FIGURE 6.29 Apples between 176 and 224 grams.

- **Solution:**
- Apples weighing 176 grams are $200-176=24$ grams below the mean, and apples weighing 224 grams are $224-200=24$ grams above the mean.
- Now 24 grams represents $24/12 = 2$ standard deviations. So the weight range of 176 grams to 224 grams is within two standard deviations of the mean.
- Therefore, about 95% of data points will lie in this range. This means that about 95% of 2000, or 1900 apples, weigh between 176 and 224 grams.

6.2 The normal distribution: Why the bell curve?

In a normal distribution, the **z-score** or **standard score** for a data point is the number of standard deviations that point lies above or below the mean. For data points above the mean the z-score is positive, and for data points below the mean the z-score is negative.

$$z - \text{score} = (\text{Data point} - \text{Mean})/\text{Standard deviation}$$

$$\text{Data point} = \text{Mean} + z - \text{score} \times \text{Standard deviation}$$

6.2 The normal distribution: Why the bell curve?

- **Example:** The weights of newborns in the United States are approximately normally distributed. The mean birthweight (for single births) is about 3332 grams (7 pounds, 5 ounces). The standard deviation is about 530 grams. Calculate the z-score for a newborn weighing 3700 grams (about 8 pounds, 2 ounces).

- **Solution:** A 3700-gram newborn is $3700 - 3332 = 368$ grams above the mean weight of 3332 grams. We divide by the number of grams in one standard deviation to find the z-score:

$$z - \text{score for } 3700 \text{ grams} = \frac{368}{530} = 0.7$$

6.2 The normal distribution: Why the bell curve?

TABLE 6.2 Percentile from *z*-Score

z-score	Percentile	*z*-score	Percentile
−2.7	0.35	0.0	50.00
−2.6	0.47	0.1	53.98
−2.5	0.62	0.2	57.93
−2.4	0.82	0.3	61.79
−2.3	1.07	0.4	65.54
−2.2	1.39	0.5	69.15
−2.1	1.79	0.6	72.57
−2.0	2.28	0.7	75.80
−1.9	2.87	0.8	78.81
−1.8	3.59	0.9	81.59
−1.7	4.46	1.0	84.13
−1.6	5.48	1.1	86.43
−1.5	6.68	1.2	88.49
−1.4	8.08	1.3	90.32
−1.3	9.68	1.4	91.92
−1.2	11.51	1.5	93.32
−1.1	13.57	1.6	94.52
−1.0	15.87	1.7	95.54
−0.9	18.41	1.8	96.41
−0.8	21.19	1.9	97.13
−0.7	24.20	2.0	97.73
−0.6	27.43	2.1	98.21
−0.5	30.85	2.2	98.61
−0.4	34.46	2.3	98.93
−0.3	38.21	2.4	99.18
−0.2	42.07	2.5	99.38
−0.1	46.02	2.6	99.53
0.0	50.00	2.7	99.65

6.2 The normal distribution: Why the bell curve?

- The **percentile** for a number relative to a list of data is the percentage of data points that are less than or equal to that number.
- **Example:** The average length of illness for flu patients in a season is normally distributed, with a mean of 8 days and standard deviation of 0.9 day. What percentage of flu patients will be ill for more than 10 days?
- **Solution:** Ten days is 2 days above the mean of 8 days. This gives a z-score of $2/0.9$ or about 2.2. Table 6.2 gives a percentile of about 98.6% for this z-score. It means that about 98.6% of patients will recover in 10 days or less. Thus, only about $100\% - 98.6\% = 1.4\%$ will be ill for more than 10 days.

6.2 The normal distribution: Why the bell curve?

- **Example:** Recall from the previous Example that the weights of newborns in the United States are approximately normally distributed. The mean birthweight (for single births) is about 3332 grams (7 pounds, 5 ounces). The standard deviation is about 530 grams.

1. What percentage of newborns weigh more than 8 pounds (3636.4 grams)?
2. Low birthweight is a medical concern. The American Medical Association defines low birthweight to be 2500 grams (5 pounds, 8 ounces) or less. What percentage of newborns are classified as low-birthweight babies?

- **Solution:**

1. $z-\text{score} = \frac{3636.4-\text{mean}}{\text{standard deviation}} = \frac{3636.4-3332}{530} = \frac{304.4}{530} = 0.6$

   Consulting Table 6.2, we find that this represents a percentile of about 72.6%.

   This means that about 72.6% of newborns weigh 8 pounds or less. So, $100\% - 72.6\% = 27.4\%$ of newborns weigh more than 8 pounds.

2. $z-\text{score} = \frac{2500-\text{mean}}{\text{standard deviation}} = \frac{3332-2500}{530} = \frac{832}{530} = 1.6$

   Table 6.2 shows a percentile of about 5.5% for a z-score of—1.6. Hence, about 5.5% of newborns are classified as low-birthweight babies.

- **The Central Limit Theorem**

According to the **Central Limit Theorem,** percentages obtained by taking many samples of the same size from a population are approximately normally distributed.

- The mean p% of the normal distribution is the mean of the whole population.
- If the sample size is n, the standard deviation of the normal distribution is:

$$\text{Standard deviation } = \sigma = \sqrt{\frac{p(100-p)}{n}} \text{ percentage points}$$

Here, p is a percentage, not a decimal.

6.2 The normal distribution: Why the bell curve?

- **Example:** For a certain disease, 30% of untreated patients can be expected to improve within a week. We observe a population of 50 patients and record the percentage who improve within a week. According to the Central Limit Theorem, the results of such a study will be approximately normally distributed.

1. Find the mean and standard deviation for this normal distribution.
2. Find the percentage of test groups of 50 patients in which more than 40% improve within a week.

- **Solution:**

1. $p = 30\%, n = 50$. A standard deviation of

$$\sigma = \sqrt{\frac{p(100-p)}{n}} = \sqrt{\frac{30(100-30)}{50}} = 6.5 \text{ percentage points}$$

2. The z-score for 40%:

$$\text{z} - \text{score} = \frac{40 - \text{mean}}{\text{standard deviation}} = \frac{40 - 30}{6.5} = \frac{10}{6.5} = 1.5$$

Table 6.2 gives a percentile of about 93.3%.

This means that in 93.3% of test groups, we expect that 40% or fewer will improve within a week.

Only $100\% - 93.3\% = 6.7\%$ of test groups will show more than 40% improving within a week.

6.2 The normal distribution: Why the bell curve?

- **Example:** Assume we know that 20% of Americans suffer from a certain type of allergy. Suppose we take a random sample of 100,000 Americans and record the percentage who suffer from this allergy.

1. The Central Limit Theorem says that percentages from such surveys will be normally distributed. What is the mean of this distribution?
2. What is the standard deviation of the normal distribution in part 1?
3. Suppose we find that in a town of 100,000 people, 21% suffer from this allergy. Is this an unusual sample? What does the answer to such a question tell us about this town?

- **Solution:**

1. The mean is $p = 20\%$.
2. For a sample size of $100{,}000$,

$$\sigma = \sqrt{\frac{p(100-p)}{n}} = \sqrt{\frac{20(100-20)}{100{,}000}} = 0.13 \text{ percentage point.}$$

3. Our sample of 21% is one percentage point larger than the mean of 20%.

$$\text{z} - \text{score} = \frac{21 - \text{mean}}{\text{standard deviation}} = \frac{21-20}{0.13} = \frac{1}{0.13} = 7.7$$

This score is far larger than any z-score in Table 6.2. There is almost no chance that in a randomly chosen sample of this size, 21% will suffer from this allergy. Thus, this is a truly anomalous sample: This town is not representative of the total population of Americans. Its allergy rate is highly unusual.

Learning Objectives:

- Understand margins of error and confidence levels in polls.
  - Basic terms: Margin of error, confidence interval, and confidence level
  - Polls: Margin of error, confidence interval, and confidence level
  - How big should the sample be?

6.3 The statistics of polling: Can we believe the polls?

- The **margin of error** of a poll expresses how close to the true result (the result for the whole population) the result of the poll can be expected to lie.
- To find the **confidence interval,** adjust the result of the poll by adding and subtracting the margin of error.
- The **confidence level** of a poll tells the percentage of such polls in which the confidence interval includes the true result.

Polls and Margin of Error

Suppose that, based on random sampling, a poll reports the percentage of the population having a certain property (e.g., planning to vote for a certain candidate) with a margin of error *m*. Assuming that this margin is based on a 95% confidence level, we can say that if we conducted this poll 100 times, then we expect about 95 of those sample results to be within *m* percentage points of the true percentage having that property.

6.3 The statistics of polling: Can we believe the polls?

- **Example:** Explain the meaning of a poll that says 33% of Americans approve of what Congress is doing, with a margin of error of 4% and confidence level of 90%

- **Solution:** In 90% of such polls, the reported approval of Congress will be within four percentage points of the true approval level.
- Thus, we can be 90% confident that the true level lies in the confidence interval between 29% and 37%.

6.3 The statistics of polling: Can we believe the polls?

Margin of Error

For a 95% level of confidence, we can estimate the margin of error when we poll *n* people using:

$$\text{Margin of error} \approx \frac{100}{\sqrt{n}}\ \%$$

Here, the symbol $\approx$ means "is approximately equal to."

6.3 The statistics of polling: Can we believe the polls?

- **Example:** A recent Oricon fashion survey asked 900 people, "Which Japanese male celebrity looks best in sneakers?" The winner was Kimura Takuya. What is the approximate margin of error for a 95% confidence level?
- **Solution:**

With n = 900:

$$\text{Margin of error} \approx \frac{100}{\sqrt{n}}\% = \frac{100}{\sqrt{900}} = 3.3\%$$

We can be 95% confident that our poll result is within 3.3 percentage points of the true value.

6.3 The statistics of polling: Can we believe the polls?

- **Example:** The Kaiser Family Foundation polled 1294 residents of Orleans Parish in New Orleans in 2008 and found that 41% of the residents who had lived through Hurricane Katrina in 2005 report that their lives are still disrupted.

1. The poll surveyed 1294 people. What is the approximate margin of error for a 95% confidence interval?
2. The poll of 1294 people found that 41% of respondents still had disrupted lives. Can we conclude with certainty that no more than 45% of residents' lives are still disrupted?
3. Suppose instead that the poll of 1294 people had found that 52% still had disrupted lives. Explain what we could conclude from this result. Could we assert with confidence that a majority of residents' lives are still disrupted by Katrina?
4. Suppose we wish to have a margin of error of two percentage points. Approximately how many people should we interview?

- **Solution:**

1. With $n = 1294$:

$$\text{Margin of error} \approx \frac{100}{\sqrt{n}} = \frac{100}{\sqrt{1294}} = 2.8\%$$

2. Our answer to part 1 tell us that we can be 95% confident that the poll number of 41% is within 2.8 percentage points of the true percentage of all residents whose lives are still disrupted from Katrina. Thus, it is very likely that the true value is:

$$\text{between } 41 - 2.8 = 38.2\% \text{ and } 41 + 2.8 = 43.8\%$$

Because the whole interval is below 45%, we can be quite confident (at a 95% level) that no more than 45% of residents' lives are still disrupted.

On the other hand, we cannot make this conclusion with absolute certainty.

- **Solution (cont.):**

3. We can be 95% confident that the poll number of 52% is within 2.8 percentage points of the true percentage of all residents whose lives are still disrupted by Katrina. Thus, it is very likely that the true value is:

$$\text{between } 52 - 2.8 = 49.2\% \text{ and } 52 + 2.8 = 54.8\%$$

Most of this interval falls above 50%, so we continue to have good reason to think that a majority of residents' lives are still disrupted.

But, because a portion of the interval falls below 50%, we should be more cautious in drawing conclusions.

- **Solution (cont.):**

4. Substitute 2% for the margin of error:

$$\text{Margin of error} = 2 = \frac{100}{\sqrt{n}} \quad \text{or} \quad \sqrt{n} = 50$$

Hence, $n = 2500$.

We should interview about 2500 people.

Note that one Harris Poll with a 95% confidence level and a margin of error of 2% surveyed 2415 people—very close to the 2500 given by the formula.

Sample Size

For a 95% level of confidence, the sample size needed to get a margin of error of *m* percentage points can be approximated using:

$$\text{Sample size} \approx \left(\frac{100}{\text{m}}\right)^2$$

- **Example:** What sample size is needed to give a margin of error of 4% with a 95% confidence level?

- **Solution:** We use the approximate formula with $m = 4$:

$$\text{Sample size} \approx \left(\frac{100}{\text{m}}\right)^2 = \left(\frac{100}{4}\right)^2 = 625$$