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?

6.1 Data summary and presentation: Boiling down the numbers

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

Chapter 6 Statistics

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.**

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.

6.1 Data summary and presentation: Boiling down the numbers

- **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:

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

Chapter 6 Statistics

6.1 Data summary and presentation: Boiling down the numbers

- **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.

Chapter 6 Statistics
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.

Chapter 6 Statistics

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

6.1 Data summary and presentation: Boiling down the numbers

- **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.

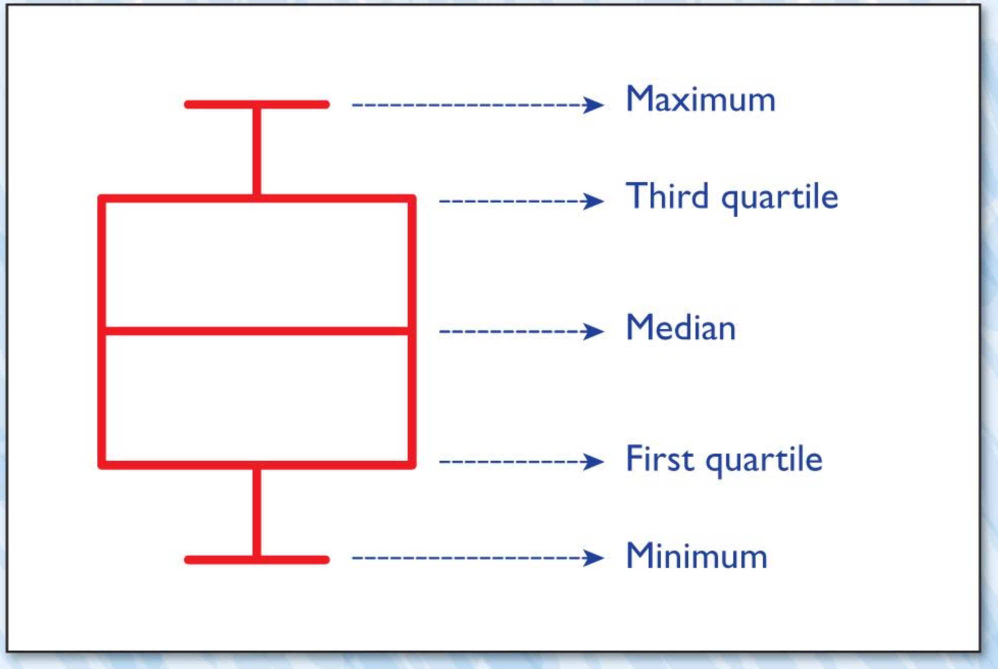

FIGURE 6.1 The basic boxplot diagram.

Chapter 6 Statistics

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.

Chapter 6 Statistics

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

6.1 Data summary and presentation: Boiling down the numbers

- **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.

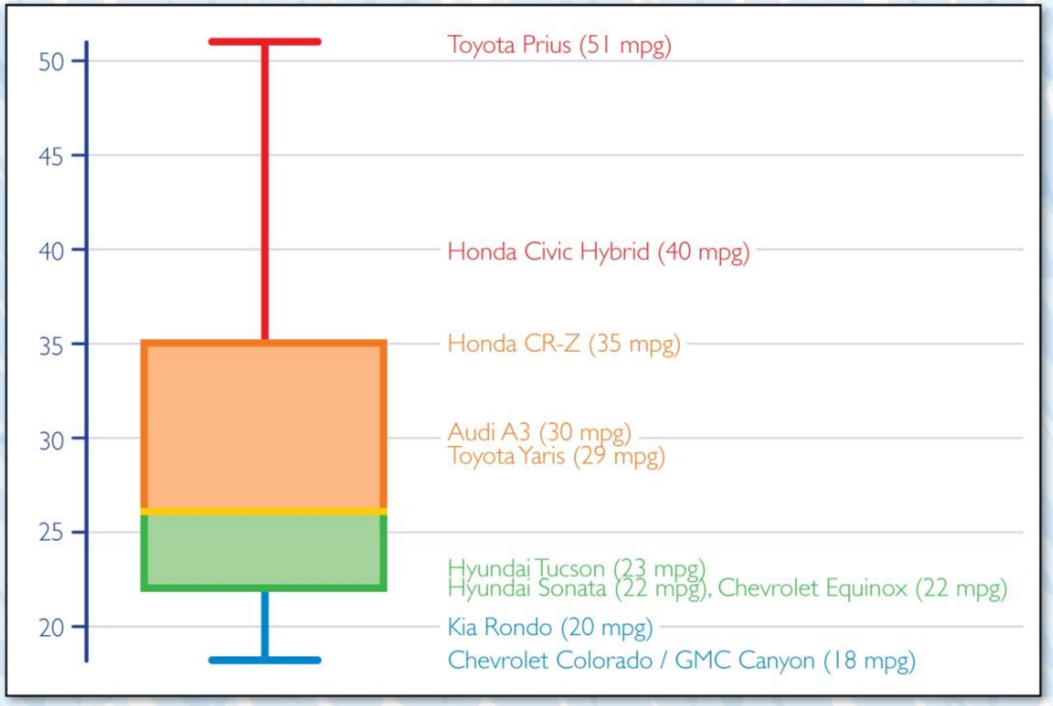

FIGURE 6.2 Boxplot for city mileage.

- **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:

$$x_1, x_2, x_3, \dots, x_n,$$

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
$\vdots$	$\vdots$	$\vdots$
x_i	$x_i - \mu$	Square of second column
$\vdots$	$\vdots$	$\vdots$
		Sum of third column Divide the above sum by *n* and take the square root.

Chapter 6 Statistics

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.

	ERA	**ERA**	**ERA**	**ERA**	**ERA**
Pitcher	**2006**	**2007**	**2008**	**2009**	**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?

6.1 Data summary and presentation: Boiling down the numbers

- **Solution:** The mean for Halladay is:

$$\mu = \frac{3.19 + 3.71 + 2.78 + 2.79 + 2.44}{5} = 2.98$$

ERA x_i	**Deviation** $x_i - 2.98$	**Square of deviation** $(x_i - 2.98)^2$
3.19	$3.19 - 2.98 = 0.21$	$(0.21)^2 = 0.044$
3.71	$3.71 - 2.98 = 0.73$	$(0.73)^2 = 0.533$
2.78	$2.78 - 2.98 = -0.20$	$(-0.20)^2 = 0.040$
2.79	$2.79 - 2.98 = -0.19$	$(-0.19)^2 = 0.036$
2.44	$2.44 - 2.98 = -0.54$	$(-0.54)^2 = 0.292$
Sum of third column		0.945
Sum divided by *n* = 5, square root		$\sigma = \sqrt{0.945/5} = 0.43$

6.1 Data summary and presentation: Boiling down the 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?

Chapter 6 Statistics

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			

6.1 Data summary and presentation: Boiling down the numbers

- **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%.

Chapter 6 Statistics

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.

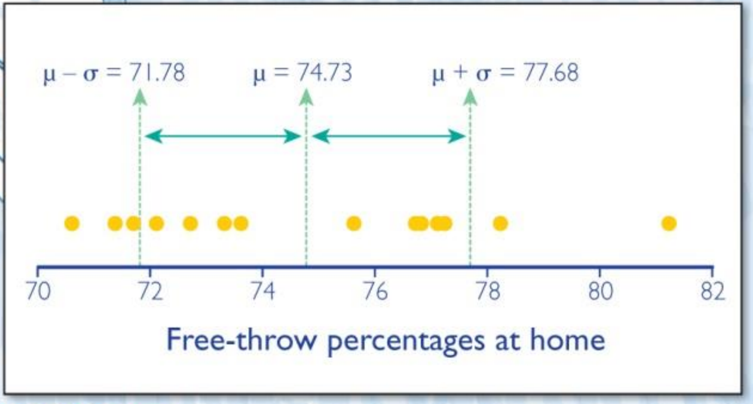

FIGURE 6.3 Eastern Conference NBA free-throw percentages at home.

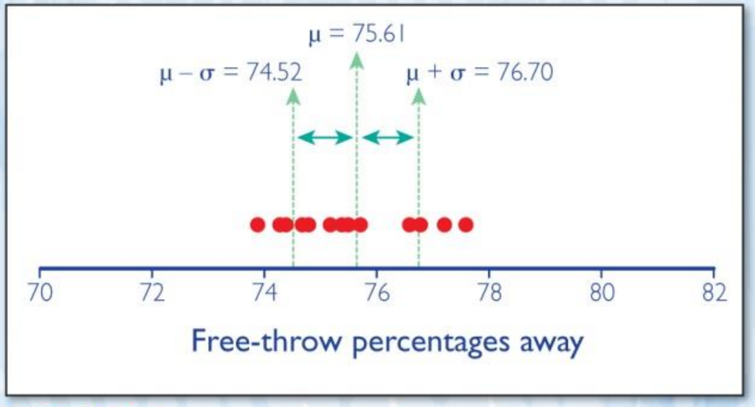

FIGURE 6.4 Eastern Conference NBA free-throw percentages away.

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.

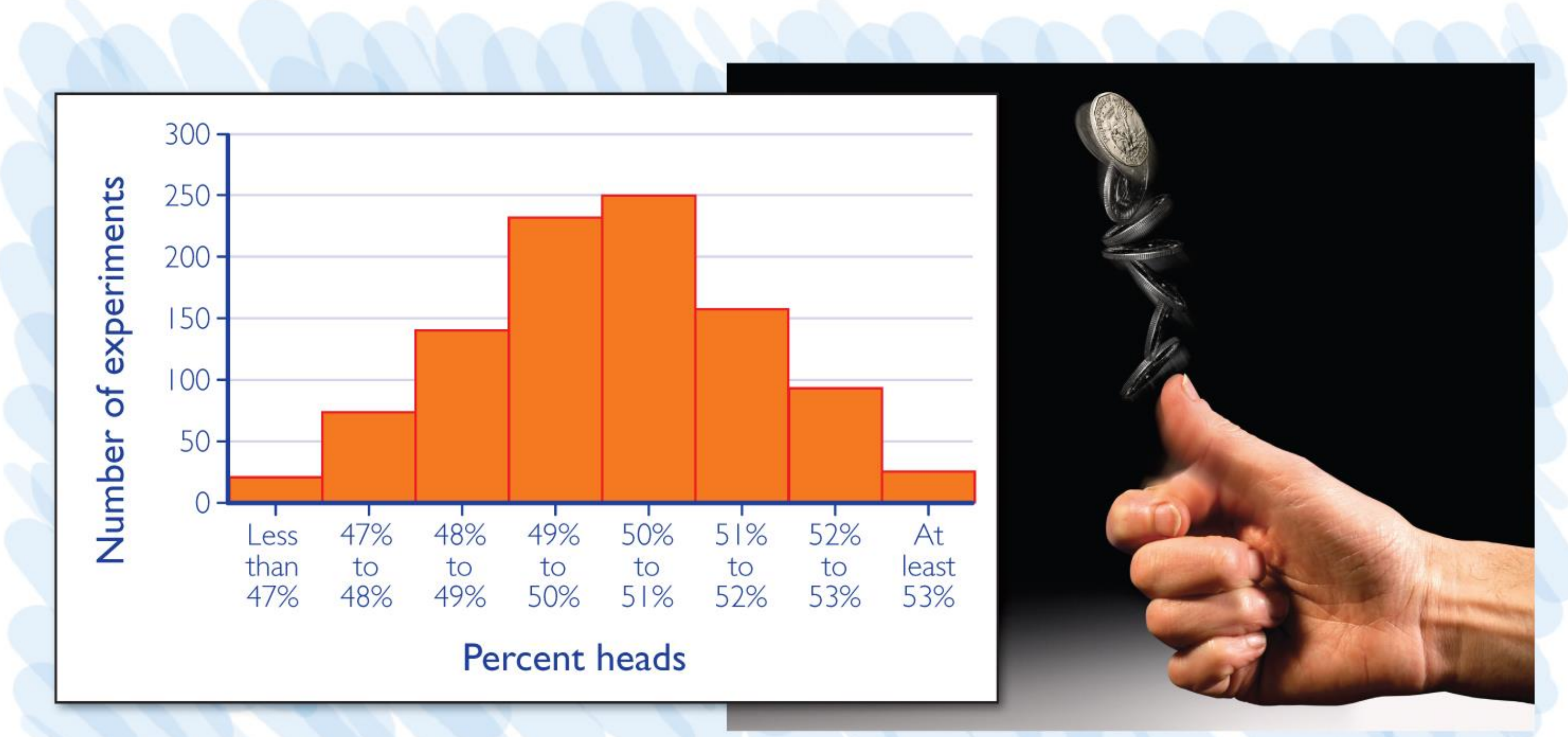

FIGURE 6.7 A histogram of coin tosses.

Chapter 6 Statistics: **Chapter Summary**

- **Data summary and presentation:** Boiling down
  - Four important measures in descriptive statistics:

    mean, median, mode, and standard deviation
- **The normal distribution:** Why the bell curve?
  - A plot of normally distributed data: the *bell-shape* curve.
  - The z-score for a data point
  - The Central Limit Theorem

Chapter 6 Statistics: **Chapter Summary**

- **The statistics of polling:** Can we believe the polls?
  - Polling involves: a margin of error, a confidence level, and a confidence interval.
- **Statistical inference and clinical trials:** Effective drugs?
  - Statistical significance and *p*-values.
  - Positive correlated, negative correlated, uncorrelated or linearly correlated