Chapter Objectives:

At the end of this chapter you should be able to:
1) create a scatterplot to graphically depict the relationship between 2 quantitative variables
2) describe the information that a scatterplot conveys about the relationship between 2 quantitative variables: form, direction, strength, points that depart from the overall pattern.
3) calculate the correlation coefficient between 2 quantitative variables using technology.
4) interpret the value of the correlation coefficient
5) describe when it is appropriate to use the correlation to describe the relationship between 2 quantitative variables
6) list the properties of the correlation coefficient
7) apply the properties of the correlation coefficient to determine the correlation when the units of the original variables are changed
8) describe the difference between association, correlation and cause-and-effect.

Motivation

1. Autism linked to precipitation levels
2. Thunderstorms linked to asthma attacks
3. Sleep disorders linked to cognitive deficits in children.
4. Osteoarthritis Risk Linked To Finger Length Ratio
5. Fast food restaurants linked to prevalence of strokes
6. Attendance at violent movies linked to reduction in crime.
How are the length of time a college student has been in school and the number of credits he/she has accumulated related?

Suppose Registration and Records at NC State tracks 100 students for 6 consecutive semesters from the Fall 2002 semester through the Spring 2005 semester. The students are all incoming freshmen (no transfer students) for the Fall 2002 semester and all 100 stay enrolled throughout the six semesters (no dropouts).

Shown below are boxplots of the cumulative credits for each semester for these 100 students.

Questions
1. Are the shapes of the distributions consistent over time?

2. Is the center consistent over time? Do we expect that given the data? Why or why not?

3. Is the variability consistent over time? Do we expect that given the data? Why or why not?

4. By examining the box plots, write down how you would explain to another person how the length of time a college student has been in school and the number of credits he/she has accumulated are related.

Scatterplots

bivariate data: \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\)

Lifespan (in years) on the y-axis and Gestation period (in days) on the x-axis for 62 different kinds of mammals.
When examining and describing a scatterplot it is sometimes easier if you separate it into several different distributions so you can determine how the centers and spread of those distributions are changing. To do this we can draw in vertical line separators for every 100 days of gestation. (You will have a vertical line at 100, 200, 300, etc.). This will give 7 slivers of data, which you can now imagine as different distributions.
1. Describe the overall **shape** of the relationship
   - Linear/curved?
   - Clusters
   - Outliers

2. Describe the **trend or direction** of the relationship

3. Describe the **strength** of the relationship
   - Strong/Weak
   - Constant/Varying

4. Are there plausible explanations for the pattern? Lurking variables?

Percent of US voters who said they would vote for a woman for US president

![Percent of US voters who said they would vote for a woman for US president](image1)

Cost per person of traffic delays in 70 US cities

![Cost per person of traffic delays in 70 US cities](image2)

To make a scatterplot: ti83/84, see p. 155; Excel, DataDesk, Appendix B, p. A-8,9
Characteristics on which to focus when examining a scatterplot

1) Form
2) Direction
3) Strength
4) Outliers

**Correlation:**

*A Quantitative Measure of the Linear Relationship Between Two Quantitative Variables*

Student weight (in pounds) vs height (in inches)

```
<table>
<thead>
<tr>
<th>Weight (lb)</th>
<th>Height (in.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>280</td>
<td>64</td>
</tr>
<tr>
<td>260</td>
<td>66</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
```

the choice of the units for the variables should not matter

Student weight (in kilograms) vs height (in centimeters)

the shape of the pattern is not changed

```
<table>
<thead>
<tr>
<th>Weight (kg)</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>120</td>
<td>160</td>
</tr>
<tr>
<td>110</td>
<td>170</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
```

Remove the units! HOW? z-scores.

standardize both variables

observation \((x_i, y_i)\) becomes
In this standardized plot:

⇒ the center of this scatterplot is at the origin
⇒ the scales on both axes are now standard deviation units
⇒ the linear pattern seems steeper since the length of one standard deviation is the same vertically and horizontally

**WARNING:** scatterplot axes can be manipulated to give a distorted visual impression of the strength of the linear relationship between x and y.

**Correlation coefficient**

\[
 r = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{n - 1}
\]

**Correlation coefficient in terms of original x and y**

\[
 r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y}
\]

The correlation coefficient measures the strength of the linear association between two quantitative variables.
DO NOT calculate correlation by hand!
To calculate correlation: using the ti83/84: see p. 151 in text; using Excel, DataDesk and other technology: see Appendix B in text, p. A-8,9.

Properties:
   i) scaleless
   ii) $-1 \leq r \leq 1$
   iii) an $r$ of $+1$ or $-1$ indicates an exact mathematical linear relationship between the $x$ and $y$ variables, that is, $y = a + bx$
   iv) correlation does not measure causation

examples:
   i) second-hand smoke
   ii) coffee & heart disease
   iii) polio in 1950's: high positive correlation between incidence of polio and soft drink consumption; both of these variables are affected by a 3rd variable: weather (covariate)

iv) many newspaper articles report evidence of correlations, but not necessarily causation.

STUDY FINDS WEIGHT GUIDELINES TOO GENEROUS

Middle-age bulge called risky

Associated Press

CHICAGO  Middle-aged women should weigh far less than people think — and less than the government recommends — in order to have healthy hearts, Harvard researchers say.

"We found that about 40 percent of all heart attacks that occur in middle-aged women are due to overweight," said Dr. JoAnn E. Manson, co-director of women's health at Harvard-affiliated Brigham and Women's Hospital in Boston. She said similar results are found in men.

The study showed that women of average weight had about a 38 percent higher risk of heart attack than women who were 15 percent less than average U. S. weights.

Women who gained 10 or fewer pounds in early to middle adulthood had the lowest risk of heart attacks, the researchers report in today's issue of The Journal of the American Medical Association.

For example, a 5-foot-4-inch woman had the lowest risk if she weighed less than 120 pounds. At the same height, a weight of 120 to 142 pounds carried a 20 percent higher risk. At 142 to 156 pounds, it was 50 percent higher; at 156 to 180 pounds it was double; and at more than 180 pounds, it was 3 1/3 times higher than for the 120 pound woman.

"I don't want to be scaring people with these findings, but we have been overly complacent about obesity and weight gain in adults," Manson said by telephone Monday.

The federal government in 1990 revised its guidelines for desirable weights upward, saying Americans over age 35 could be significantly heavier than under 1965 guidelines.

"The current federal weight guidelines are in a sense encouraging the fattening of Americans," Manson said, noting that one in three adults is overweight.

While cautioning against overreaction to the findings, she recommended increasing physical activity, lowering the fat and calorie content of the diet and eating more fruits, vegetables, and grains.

Important: before you use the correlation to describe the realtionship between 2 variables, check the following conditions:

1) Quantitative variables condition.
   correlation applies only to quantitative variables; don't apply correlation to categorical variables masquerading as quantitative variables

2) Straight enough condition
   correlation measures the strength only of the linear association, and will be misleading if the relationship is not linear. What is “straight enough”? That's a judgment call.

3) Outlier condition.
   Outliers can dramatically distort the correlation. If there is an outlier, report the correlation both with and without the outlier.

Examples:
   1) Car weight and fuel consumption
      $x =$ car weight (in thousands of pounds)
      $y =$ fuel consumption (gallons needed to go 100 miles)
2) SAT scores and percentage of hs seniors taking the test
x = percentage of high school seniors in a state taking the test
y = mean SAT score in a state

Properties of the correlation:

1) scaleless (no units)
2) \(-1 \leq r \leq 1\)
   - \(r = -1\) only if \(y = a + bx\) with slope \(b<0\)
   - \(r = +1\) only if \(y = a + bx\) with slope \(b>0\)
3) It doesn’t make any difference which variable you call “x” and which variable you call “y”
4) \(r\) is not affected by a linear transformation on \(x\) or \(y\) if the slope \(b > 0\).
Example (linear transformation does not affect correlation)

temperature is in °F

![Mean Annual Central Park Temp vs Year](image)

![Mean Annual Central Park Temp (C) vs Year](image)

What can Go Wrong

1. Don't say “correlation” when you mean “association”

   **Association** is a deliberately vague term.  
   **Correlation** is a precise term describing the strength and direction of the **linear** relationship between 2 quantitative variables.

2. Don't correlate categorical variables

   The correlation between gender of worker and salary is 0.75.

3. Be sure the association is linear

   Plot the data!  
   In the graph below the small correlation could give the impression that there is no relationship between speed and miles per gallon.  
   The relationship is **nonlinear**.

5. Don't confuse correlation with causation.