Chapter 10

Two Quantitative Variables

Chapter Overview

In Chapters 5–9 the focus was on comparing two or more groups. Chapter 10 is different. Until now the explanatory variable has been categorical. In Chapter 10 both variables are quantitative. This allows us to address questions about an association between two quantitative variables and to predict the value of one variable based on the other variable.

Graphical summaries will use a new kind of plot called a scatterplot. These are like the dotplots we’ve been using throughout the book, except that they now tell you information about two (quantitative) variables at once. In describing these graphs, we will focus on the Form, Direction, and Strength of the relationship. Here are some quick examples: (1) You measure the heights of classmates, first in centimeters, then in inches. Graphing the height in centimeters on the y-axis vs. the height in inches along the x-axis, the dots will fall along a straight line. The direction of the relationship is positive: y goes up as x goes up. The relationship cannot be stronger, because the points lie exactly on the line and we can perfectly predict one variable from the other. (2) You measure shoe size and height of your classmates. Taller people tend to have bigger feet (the relationship is positive), but there is a lot of variability (the relationship is not as strong). (3) You plot age and height for one person each year on her birthday, from year 1 to year 60. This relationship is not linear – people stop getting taller, and the plot flattens out. (4) You plot age and height for a sample of adults. There may be no systematic relationship. Once we stop getting taller, age and height for a sample of adults are unrelated.

Chapter 10 has five sections. The first section uses scatterplots to explore form, direction, and strength of a relationship between two quantitative variables, and introduces a new statistic, the correlation coefficient, which measures direction and strength of a linear association on a scale from -1 to 1, with 0 indicating no linear relationship. (Heads up: The strength of a relationship is not the same as the strength of evidence that a relationship exists.) Section 10.2 illustrates how to conduct simulation-based tests based on the correlation coefficient.

Section 10.3 indicates how to fit a line to give a quantitative description of a linear relationship. The slope of the fitted line might tell us, for example, that between the ages of 8 and 13, kids tend to get taller by
about 2 inches per year. Section 10.4 uses simulation to test a null hypothesis of no association using the slope as the statistic. Finally, Section 10.5 presents a theory-based approach for both the regression slope and the correlation coefficient.

It is important to keep in mind throughout this chapter that association does not imply causation. Beware of confounding variables. For a sample of children in grades K-12, kids with larger feet tend to drink more diet soft drinks, but drinking diet soda won’t make your feet grow. State by state, there is a strong positive relationship between the number of McDonald’s restaurants and the number of cases of lung cancer, but Big Macs do not cause lung cancer.

**Section 10.1: Summarizing the Relationship Between Two Quantitative Variables Using the Correlation Coefficient**

*Introduction*

Do you think there is an association between a student’s grade point average and the number of classes they skip per semester? How about the fuel efficiency of cars and their weight? These questions are different from what we have studied so far, because both variables are *quantitative*. But we will follow the same general strategy as before – first producing graphical summaries that can illuminate the relationship exhibited in the sample and then numerical summaries or statistics that measure the strength of the association. In the next section (10.2), we will draw inferences, again using simulation to compare the observed statistic of association to what we might expect to see by chance alone. These inferential techniques employ the same reasoning process as techniques we used in previous chapters. We start, however, by spending some time with scatterplots and the correlation coefficient before moving into simulation-based inference techniques.

**Example 10.1: Exam Times and Exam Scores**

Is there an association between the time it takes a student to take an exam and their score on the exam? One of the authors wondered about this question and decided to collect data from one of his mathematics classes.
**Think about it:** Which variables do you consider to be explanatory and response?

Even though this was not an experiment, it makes sense to consider the time spent on the test as the explanatory variable, helping to explain why some scores were higher and some were lower. It also makes sense to think of predicting test score (response) from time spent on test (explanatory). Note: With this goal, we often call the explanatory variable a *predictor* variable.

**Exploring the Data: Graphical Summary**

To examine an association between these variables graphically, it would not be helpful to make separate dotplots of time spent on test and score on test. Doing so would do nothing to tell us about the *relationship* between the two variables. Instead, a *scatterplot* shows the two values for each observational unit. In constructing a scatterplot we put the explanatory variable on the horizontal axis and the response variable on the vertical axis. (If you don’t have a clear explanatory/response variable distinction, then this choice is more arbitrary.) Then each dot on the scatterplot shows the values for both variables on that observational unit. The results for 22 students are shown in Figure 10.1.

**Figure 10.1:** A scatterplot showing the association between students’ scores on a mathematics test and the time it took them to take the test

**Think about it:** Based on the graph, approximate the score and time for the student who was the first to turn in her/his exam. Also describe what you learn about these data from the scatterplot.
When describing a scatterplot we look for three aspects of association: direction, form, and strength. The *direction* of the association in Figure 10.1 is negative because larger values on one variable correspond to smaller values on the other variable. This shows that the longer it takes students to finish the test, the lower their scores tend to be. (This is not a hard and fast rule of course, some student who took longer scored higher; but overall, there is this tendency.)

In describing the *form* of an association, we describe whether it follows a linear pattern or some more complicated curve. For the test time and score data, it appears that a line would do a reasonable job of summarizing the overall pattern in the data, so we say that the association between score and time is linear. We will see lots of examples in which the form is fairly linear. Occasionally, however, the form is clearly not linear and that will be very important to note.

In describing the *strength* of association revealed in a scatterplot, we see how closely the points follow a pattern. If you imagine a line through the test time and score scatterplot, the data points all fall pretty close to that line, so we say that the association is reasonably strong.

It is also important to consider any *unusual observations* that do not follow the overall pattern. In this example, the fastest student finished in about 30 minutes and scored 100 on the exam. Even though this time is smaller than all the rest, and might be considered an outlier if looking at the time variable alone (no one else finished in less than 40 minutes!), this (time, score) combination appears consistent with the overall pattern of the sample. However, if we saw a student that finished in 30 minutes and only received a score of 60, we might try to follow-up on the data and make sure there wasn’t an error recording their score, their time, or some other issue.

**Key idea:** Association between quantitative variables can be described with direction, form, and strength.

- If above average values of one variable tend to correspond to above average values of the other variable, the direction is positive. If, however, above average values of one variable are associated with below average values of the other, the direction is negative.
- The form of the association is whether the data follow some linear pattern or some more complicated pattern.
- The strength of an association refers to how closely the data follow a particular pattern. When an association is strong, the value of one variable can be accurately predicted from the other.
- Also look for and investigate further any unusual observations that do not fit the pattern of the rest.

**Exploring the Data: Numerical Summary**
We can also summarize the association between two quantitative variables numerically. The *correlation coefficient* measures the strength and the direction of a *linear* association between two quantitative variables. The correlation coefficient, usually denoted by the letter $r$, has the property that $-1 \leq r \leq 1$. If the correlation coefficient is close to -1 or 1, then the variables are highly correlated, revealing a strong linear association. In fact, if $r = 1$ or $r = -1$, the data all fall exactly in a straight line. If the correlation coefficient is close to 0, then the data points do not have a linear association. If the correlation coefficient is positive, then there is a positive linear association between the two variables. If the correlation coefficient is negative, then there is a negative linear association between the two variables.

**Key Idea:** The correlation coefficient uses a rather complex formula (see the Appendix Calculation Details for details) that is rarely computed by hand; instead, people almost always use a calculator or computer to calculate the value of the correlation coefficient. But you should be able to apply the above properties to interpret the correlation coefficient that is found.

The correlation coefficient between test time and score is approximately $r = -0.56$. The negative value indicates that students with longer times tend to have lower test scores. The magnitude of the -0.56 correlation appears to indicate a fairly strong linear association between time and test score, so you could predict a student’s test score reasonably well knowing only how long he/she took on the test. Be aware that evaluating the strength of the association will be subjective, especially across different research areas. A correlation of -0.56 might be considered extremely strong for a psychologist trying to predict human behavior but very weak for a physicist trying to predict how far a particle will travel. (To help develop your judgment for how a scatterplot looks for various correlation values, you can play the Correlation Guessing Game applet.)

**Key Idea:** The correlation coefficient is only applicable for data which has a linear form; non-linear data is not summarized well by the correlation coefficient. In fact, we could say that the correlation coefficient is a numerical summary of the strength and direction of a linear association between two quantitative variables.

**Caution: Influential Observations**

We haven’t been completely honest in presenting our test time and score dataset. There were three additional students in the class, the last three students to turn in their tests. Their times were 93, 96, and 100 minutes. These three students also scored very well on the exam: 93, 93, and 97 respectively. We added these scores to the scatterplot shown in Figure 10.2.
**Figure 10.2:** A scatterplot of the score on a math test and the time (in minutes) it took to complete the test after the last three tests turned in were added for all 25 students.

The three added scores can be seen on the far right side of the graph. These three students took a long time and had high scores, so they go against the general tendency that students who took longer tended to score lower. Without these three students we would view the relationship differently than we do with these students. For this reason, these unusual observations are called *influential observations*. They are influential because the addition of just these three points changes the correlation coefficient dramatically (and our perception of the strength of the relationship), from -0.56 (a fairly strong negative correlation) to -0.12 (a weak negative correlation). Influential points tend to be ones that are far to the right or left on the graph, having extreme values for the explanatory variable like these three. The amount of influence of even just one observation can sometimes be dramatic. For example, if the person that finished the test in 30 minutes would have scored 50 instead of 100, the correlation coefficient would switch from negative to positive (0.13).

Note: There are two kinds of unusual observations seen in scatterplots: influential observations and outliers. An observation is *influential* if removing it from the data set dramatically changes our perception of the association. Typically influential observations are extreme in the explanatory variable. Outliers are observations that don’t follow the overall pattern of the relationship. They may or may not be influential and they may or may not be extreme in either variable individually, but are unusual in terms of the combination of values.

**Exploration 10.1: Are Dinner Plates Getting Larger?**
Figure 10.3: Are the squares the same size? Or is one larger than the other?

1. Look at the two pictures in Figure 10.3. What do you think: Are the squares the same size, or is one larger than the other?

For many people, the square on the right looks smaller, but in fact they are about the same size.

You have probably seen many articles and television reports about the growing obesity problem in America. Many reasons are given for this problem, from fast food to lack of exercise. One reason that some have given is simply that the size of dinner plates are increasing. Figure 10.3 illustrates that that people may tend to put more food on larger dinner plates without even knowing they are doing so.

Step 1: Ask a research question
An interesting research question is whether dinner plates are getting larger over time.

Step 2: Design a study and collect data.

To investigate the claim that dinner plates are getting larger, some researchers gathered data for American dinner plates being sold on ebay.com on March 30, 2010 (Van Ittersum and Wansink, 2011). They recorded the size and the year of manufacture for each distinct plate in a sample of 20 American-made dinner plates.

2. What is the explanatory variable? Is it categorical or quantitative?

3. What is the response variable? Is it categorical or quantitative?
4. In your own words, explain what it would mean for there to be an association between the explanatory and response variables in this study.

5. What are the observational units? How many are there?

6. Is this study an experiment or an observational study? Why?

This study is different than studies we’ve looked at before because both of the variables of interest are quantitative. For this reason, the graphical and numerical techniques we will use to summarize the data are different than what we’ve seen before.

Step 3: Explore the data.

The data in Table 10.1 represent a subset of the data values reported in the research paper.

Table 10.1: Data for size (diameter, in inches) and year of manufacture for 20 American-made dinner plates

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>size</td>
<td>10</td>
<td>10.75</td>
<td>10.125</td>
<td>10</td>
<td>10.625</td>
<td>10.75</td>
<td>10.625</td>
<td>10</td>
<td>10.5</td>
<td>10.125</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>size</td>
<td>10.375</td>
<td>10.75</td>
<td>10.375</td>
<td>11</td>
<td>10.75</td>
<td>10.125</td>
<td>11.5</td>
<td>11</td>
<td>11.125</td>
<td>11</td>
</tr>
</tbody>
</table>

Graphical summary of two-quantitative variables: Scatterplots

7. Paste the PlateSize data set into the Corr/Regression applet. Which variable is on the x (horizontal) axis? Which variable is on the y-(vertical) axis?
A *scatterplot* is a graph showing a dot for each observational unit, where the location of the dot indicates the values of the observational unit for both the explanatory and response variables. Typically, the explanatory variable is placed on the \(x\)-axis and the response variable is placed on the \(y\)-axis.

8. Create a rough sketch of the scatterplot from the applet here. Be sure that the axes are labeled well. Show that you understand how a scatterplot is created by circling the dot on the scatterplot corresponding to a plate made in 1958 that is 10 inches in diameter.

When describing a scatterplot we look for three aspects of association: direction, form, and strength.

The *direction* of association between two quantitative variables is either positive or negative, depending on whether or not as the explanatory variable (year) increases the response variable (size) tends to increase (positive association) or decrease (negative association).

9. **Direction.** Is the association between year and plate size positive or negative? Interpret what this means in context.

The *form* of association between two quantitative variables is described by indicating whether a line would do a reasonable job summarizing the overall pattern in the data or if a curve would be better. It is important to note that, especially when the sample size is small, you don’t want to let one or two points on the scatterplot to change your interpretation of whether or not the form of association is linear. In general, assume that the form is linear unless there is compelling (strong) evidence in the scatterplot that the form is not linear.
10. **Form.** Does the association between year and size appear to be linear, or is there strong evidence (many observational units) suggesting the relationship is non-linear?

In describing the *strength* of association revealed in a scatterplot, we see how closely the points follow the form: that is how closely do the points follow a straight line or curve. If all of the points fall pretty close to this straight line or curve, we say the association is strong. Weak associations will show little pattern in the scatterplot, and moderate associations will be somewhere in the middle.

11. In your opinion, would you say that the association between plate size and year appears to be strong, moderate or weak?

It is also important to consider any *unusual observations* that do not follow the overall pattern.

12. Are there any observational units (dots on the scatterplot, representing individual plates) that seem to fall outside of the overall pattern?

In this example, the largest plate diameter was 11.5 inches. Compared to other plate diameters in the sample, this is fairly large. But based on the scatterplot we see that the plate doesn’t appear to be unusual. On the other hand, if the 11.5 inch plate had been manufactured in 1950, then the scatterplot would indicate that to be unusual.

Note: There are two kinds of unusual observations seen in scatterplots: influential observations and outliers. An observation is *influential* if removing it from the data set dramatically changes our perception of the association. Typically influential observations are extreme in the explanatory variable. *Outliers* are observations that don’t follow the overall pattern of the relationship. They may or may not be influential and they may or may not be extreme in either variable individually, but are unusual in terms of the combination of values.
Key idea: Association between quantitative variables can be described with direction, form, and strength.

- If above average values of one variable tend to correspond to above average values of the other variable, the direction is positive. If, however, above average values of one variable are associated with below average values of the other, the direction is negative.
- The form of the association is whether the data follow some linear pattern or some more complicated pattern.
- The strength of an association refers to how closely the data follow a particular pattern. When an association is strong, the value of one variable can be accurately predicted from the other.
- Also look for and investigate further any unusual observations that do not fit the pattern of the rest.

Numerical Summaries

Describing the direction, form, and strength of association based on a scatterplot, along with investigating unusual observations, is an important first step in summarizing the relationship between two quantitative variables. Another approach is to use a statistic. One of the statistics most commonly used for this purpose is the correlation coefficient.

The correlation coefficient, often denoted by the symbol $r$, is a single number that takes a value between -1 and 1, inclusive. Negative values of $r$ indicate a negative association, whereas positive values of $r$ indicate a positive association.

13. Will the value of the correlation coefficient for the year-plate size data be negative or positive? Why?

Key Idea: The correlation coefficient uses a rather complex formula (see the Appendix Calculation Details for details) that is rarely computed by hand; instead, people almost always use a calculator or computer to calculate the value of the correlation coefficient. But you should be able to apply the above properties to interpret the correlation coefficient that is found.

The stronger the linear association is between the two variables, the closer the value of the correlation coefficient will be to either -1 or 1, whereas linear associations will have correlation coefficient values closer to 0. Moderate linear associations will typically have correlation coefficients in the range of 0.3 to 0.7, or -0.3 to -0.7.
14. Without using the applet, give an estimated range for the value of the correlation coefficient between plate size and year based on the scatterplot:

15. Now, check the **Correlation coefficient** box in the applet to reveal the actual value of the correlation coefficient. Report the value here.

**Key Idea:** The correlation coefficient is only applicable for data which has a linear form; non-linear data is not summarized well by the correlation coefficient. In fact, we could say that the correlation coefficient is a numerical summary of the strength and direction of a linear association between two quantitative variables.

Another point about correlation is that it is sensitive to unusual observations called influential points. Influential points are unusual observations which substantially impact the value of the correlation based on whether or not they are present in the dataset.

16. In the **Add/remove observations** section of the applet, enter a year \( x \) of 1950 and plate size \( y \) of 11.5 and press **Add**. Note how this is an unusual observation on the scatterplot. How did the correlation change as a result?

We have now seen a graphical technique (scatterplot) and numerical summary (correlation coefficient) to summarize the relationship between two quantitative variables. Using techniques you will learn in the next section, this correlation coefficient is significantly different than zero (p-value \( \approx 0.01 \)), meaning that there is strong evidence of an association between year and plate size, though we need to be cautious about generalizing too far because this is only a random sample of plates for sale on a single day on eBay.

**Section 10.1 Summary**

Examining the **association** between two quantitative variables involves looking at scatterplots.
• The direction of the association indicates whether large values of one variable tend to go with large values of the other (positive association) or with small values of the other (negative association).

• The form of association refers to whether the relationship between the variables can be summarized well with a straight line or some more complicated pattern.

• The strength of association entails how closely the points fall to a recognizable form such as a line.

• As always, unusual observations that do not fit the pattern of the rest of the observations are worth examining closely or have an undue impact on our perception of the association.
  o Influential observations are observations whose removal would dramatically impact the overall direction, form, or strength of the association. (Often you find them by looking for observations that are extreme in the x-direction.)
  o Outliers are unusual observations that do not follow the overall pattern of the association. They may or may not be influential.

A numerical measure of the association between two quantitative variables is the correlation coefficient, \( r \).

• A correlation coefficient is between -1 and 1, inclusive.
• The closer the correlation coefficient to -1 or 1, the stronger the linear association between the variables.
• The closer the correlation coefficient to 0, the weaker the linear association between the variables.
• Correlation coefficient is not resistant to extreme values.

Section 10.2: Inference for the Correlation Coefficient: A Simulation-based Approach

Introduction

In the previous section we learned about how the correlation coefficient and scatterplots are used to describe the association between two quantitative variables for the sample. We will now learn how to test whether there is convincing evidence of any association between the two quantitative variables in the larger population or process.

Example 10.2: Exercise Intensity and Mood Changes
You have probably heard that exercise helps relieve stress. Exercising has been shown to decrease tension and elevate one’s mood. However, is this change in mood related to the intensity of the exercise? Szabo (2003) investigated this question with a sample of female college students who jogged or ran on a treadmill for 20 minutes. Participants were allowed to go at any speed they desired as long as they didn’t walk. Using heart rate monitors, researchers calculated their exercise intensity as a percentage of their maximal heart rate. (The subjects’ maximal heart rates were estimated from their base heart rates. As a result a couple of these percentages are over 100.)

Before exercising, the subjects also took a quick test about their mood. The test measured things like anger, confusion, and depression. This test was given again after exercising, and the researchers determined a numerical score for the change in mood of the subjects before and after exercising. Negative numbers indicate a decrease in measures of bad mood like anger, and so forth. Although the researchers found significant results in the participants’ change of mood when comparing before and after exercise, we are interested to see whether there is a significant association between the change of mood and exercise intensity. Let’s look at this study using the six step statistical investigation method.

**Step 1: Ask a research question.** Is there an association between how much the mood of a person changes during exercise and the intensity of the exercise?

**Step 2: Design a study and collect data.** The study design was detailed above. What we haven’t done yet, however, is to convert our research question into hypotheses. Writing our hypotheses out in words should look very familiar to the way we have done this in the past few chapters.

Null Hypothesis: There is no association between exercise intensity and changes in mood in the population.

Alternative Hypothesis: There is an association between exercise intensity and changes in mood in the population.

We are going to use the correlation coefficient to measure the association between these two variables. See FAQ 10.2.1 for further discussion.

The researchers used 32 volunteer female college students as their subjects. Their results are shown in Table 10.2.
**Table 10.2:** Exercise intensity as a percentage of maximal heart rate and the change in mood test scores

<table>
<thead>
<tr>
<th>Ex. Intensity</th>
<th>MoodChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>41.8</td>
<td>-12</td>
</tr>
<tr>
<td>56.1</td>
<td>-4</td>
</tr>
<tr>
<td>56.8</td>
<td>-5</td>
</tr>
<tr>
<td>59.1</td>
<td>-6</td>
</tr>
<tr>
<td>59.5</td>
<td>-24</td>
</tr>
<tr>
<td>60.9</td>
<td>1</td>
</tr>
<tr>
<td>61.6</td>
<td>-24</td>
</tr>
<tr>
<td>62.6</td>
<td>-3</td>
</tr>
<tr>
<td>64.9</td>
<td>-5</td>
</tr>
<tr>
<td>65.6</td>
<td>-12</td>
</tr>
<tr>
<td>66.4</td>
<td>-16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ex. Intensity</th>
<th>MoodChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>66.5</td>
<td>-3</td>
</tr>
<tr>
<td>67.1</td>
<td>-8</td>
</tr>
<tr>
<td>67.8</td>
<td>-2</td>
</tr>
<tr>
<td>68.3</td>
<td>-24</td>
</tr>
<tr>
<td>68.4</td>
<td>-10</td>
</tr>
<tr>
<td>69.7</td>
<td>-12</td>
</tr>
<tr>
<td>69.9</td>
<td>-4</td>
</tr>
<tr>
<td>70.1</td>
<td>-12</td>
</tr>
<tr>
<td>70.3</td>
<td>-18</td>
</tr>
<tr>
<td>72</td>
<td>-13</td>
</tr>
<tr>
<td>72.2</td>
<td>-31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ex. Intensity</th>
<th>MoodChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>73.9</td>
<td>-20</td>
</tr>
<tr>
<td>76.1</td>
<td>-26</td>
</tr>
<tr>
<td>77</td>
<td>-3</td>
</tr>
<tr>
<td>77.3</td>
<td>8</td>
</tr>
<tr>
<td>79.9</td>
<td>2</td>
</tr>
<tr>
<td>82.5</td>
<td>-11</td>
</tr>
<tr>
<td>83.3</td>
<td>-3</td>
</tr>
<tr>
<td>87</td>
<td>5</td>
</tr>
<tr>
<td>106</td>
<td>-13</td>
</tr>
<tr>
<td>109.6</td>
<td></td>
</tr>
</tbody>
</table>

**Step 3: Explore the data.** A scatterplot of the change in mood and exercise intensity is shown in Figure 10.4.

**Figure 10.4:** The data were put in the Corr/Regression applet and a scatterplot of change in mood and exercise intensity was produced.

Think about it: Do these data support the research conjecture?

The scatterplot shows a weak positive association between change in mood and intensity of exercise, with a sample correlation of $r = 0.187$. 
Now suppose there is no association between these two variables in the population. Would it be *possible* to get a sample correlation coefficient as large as $r = 0.187$ just by random chance? The answer to this question, as always, is yes, it’s possible. The real question though, is whether it is *unlikely*. Answering this question will take more investigation.

One way to explore this question is to investigate how much sample to sample variability we can expect to see in the correlation coefficient when the two variables are not related; that is, if the change in mood numbers were essentially assigned at random to the exercise intensity numbers, how strong is the typical “by chance” association?

**Think about it:** Suggest a procedure for evaluating the strength of evidence against the null hypothesis.

**Step 4: Draw inferences.** We will proceed just as before with our simulation-based approach, starting with the 3S Strategy.

1. **Statistic:** In this case, we found the observed correlation coefficient between exercise intensity and mood change for our sample to be $r = 0.187$.

2. **Simulate:** We will assume there is no relationship between change in mood and exercise intensity in the population (this is our null hypothesis). If this is the case, it does not matter which change in mood value paired up with which exercise intensity value. Therefore we will shuffle the response variable, change in mood, and find the correlation coefficient for our new shuffled data. We will repeat this random shuffling a large number of times to obtain a null distribution that models the behavior of the correlation coefficient under the null hypothesis.

To better understand the simulation approach, consider the following copy of the original data table where we have removed the mood change values from the table.
Imagine that the actual mood change scores were now randomly matched to the different exercise intensity values. You can envision this random matching process by imagining that the 32 different mood change scores were each written on a slip of paper, so you could shuffle all the mood change scores and randomly pair up each mood change score with an exercise intensity score. Thus, this simulation means that for this process there really is no underlying association between exercise intensity and mood change, any association observed in the sample is simply due to chance.

3. **Strength of evidence:** If the observed statistic (correlation coefficient of our sample, \( r = 0.187 \)) is in the tail of the null distribution created by our simulation, then we say we have found evidence against the null hypothesis in favor of the alternative hypothesis.

To complete steps 2 and 3 of our 3S strategy, we used the Corr/Regression applet to shuffle the mood change values 1000 times, and after each shuffle calculated the correlation coefficient. These 1000 shuffled correlation coefficients are shown in a histogram in Figure 10.5. We counted that 328 of these 1000, or 32.8%, were as large as or larger than our sample correlation coefficient of 0.187 or as small or smaller than -0.187 (two-sided), giving us an approximate p-value of 0.328. Therefore, our sample correlation coefficient is not very unlikely to occur by chance alone when the null hypothesis is true. Hence these data do not provide strong evidence against the null hypothesis and do not support the alternative hypothesis that change in mood is associated with intensity of exercise in the population.
Figure 10.5: The null distribution of correlation coefficients for 1000 shuffles of mood change assuming no association between mood change and exercise intensity

Here is a summary of the simulation model:

- **Null hypothesis** = No association between $x$ and $y$ variables
- **One repetition** = Re-randomizing the response outcomes to the explanatory variable outcomes
- **Statistic** = correlation coefficient

**Step 5: Formulate conclusions.** We do not have strong evidence that there is a genuine association between change in mood and exercise intensity in the population. Hence, we cannot really talk about cause-and-effect. Even if we found a significant association, we would still not be able to draw a cause-and-effect conclusion because this was an observational study in which subjects chose for themselves the level of exercise intensity. Also, this study does not appear to have selected a random sample, so we should be cautious in generalizing these results to a larger population. At best, they apply to female college students. The study did report that they had significant results when comparing mood before and after exercise, however the intensity of the exercise does not seem to significantly influence the amount of the change in mood for students like those in this study.
Step 6: Look back and ahead. The researchers in this study now have some important decisions to make. The results of this analysis did not provide compelling evidence that exercise intensity and mood change were associated in the population. But is this an example of a missed opportunity to establish that such a hypothesis is true? Or, are the researchers incorrect in what they’ve hypothesized? Although we’ll never know for sure from a single study, there are a number of possible improvements if the researchers do a follow-up study: including potentially
(a) gathering a larger sample size in a follow-up study,
(b) considering whether that sample should be randomly chosen from a population of interest, (c) randomly assigning subjects to different exercise intensity levels,
(d) improving the way that “mood” and “mood change” were measured (e.g., using reliable and valid instruments),
(e) considering starting only with subjects who had a negative mood before exercising as those with positive moods before exercising may show little change afterwards.

FAQ 10.2.1 Where’s the parameter? Why no interval?

Q: I’ve got a question about the draft lottery (Exploration 10.2).
A: OK.
Q: What’s the parameter?
A: That’s a complicated question. The short answer is there isn’t one, but to explain why, I want to go back and get some background. Are we asking about a process or a population?
Q: A process. Mix the numbered capsules, then take them out one at a time.
A: Right. And the null hypothesis?
Q: There’s no association between sequential date and draft number.
A: Right. Which number goes with which birthday is purely a matter of random chance. What’s the alternative hypothesis?
Q: There’s some sort of association. The pattern in the scatter plot is not due just to random chance.
A: Right.
Q: I’m still waiting. What’s the parameter?
A: One more question. Suppose the null is false and there is some sort of association. What form will the association take?
Q: How should I know? There could be all sorts of patterns hidden in that scatterplot. Possibly
linear, possibly curved, possibly up and down. Who knows.
A: Exactly. Until we can say more about the structure of the alternative – about the form of the pattern – we can’t specify a parameter.
Q: You took long enough to say you don’t know.
A: Yes, but there’s an important lesson: You have to know (or assume) some important features of your data in order have a meaningful parameter. In early chapters we assumed the probability of success or the difference in treatment means was all that mattered, so those were our parameters.
Q: But you can measure strength of evidence against the null even without a parameter? Seems like magic.
A: No magic. As always, there’s no free lunch. You don’t need a parameter, but your test may not be very powerful, because it has to work for all possible alternatives.
Q: I thought I asked a simple question.
A: An important question, even if the answer isn’t simple.
Q: One last question. Why don’t we do simulation-based confidence intervals?
A: The short answer is that you can’t have confidence intervals if you don’t have a parameter. But we’ll come back to all of this in Section 5.
Q: I can hardly wait. Be still my heart!
**Exploration 10.2: Draft Lottery**

In 1970, the United States Selective Service conducted a lottery to decide which young men would be drafted into the armed forces (Fienberg, 1971). Each of the 366 birthdays in a year (including February 29) was assigned a draft number. Young men born on days assigned low draft numbers were drafted.

We will regard the 366 dates of the year as observational units. We will consider two variables recorded on each date: *draft number* assigned to the date, and *sequential date* in the year (so January 31 is sequential date 31, February 1 is sequential date 32, and so on).

1. In a perfectly fair, random lottery, what should be the value of the correlation coefficient between *draft number* and *sequential date of birthday*?

Figure 10.6 displays a *scatterplot* of the assigned draft numbers and the sequential dates. There are 366 dots, one for each day of the (leap) year.

**Figure 10.6:** A scatterplot of the draft numbers and sequential dates

2. Let's take a look at the scatterplot in Figure 10.6.
   a. Does the scatterplot reveal much of an association between draft number and sequential date?
   b. Based on the scatterplot, guess the value of the correlation coefficient.
c. Does it appear that this was a fair, random lottery?

It’s difficult to see much of a pattern or association in the scatterplot, so it seems reasonable to conclude that this was a fair, random lottery with a correlation coefficient near zero.

But let’s dig a little deeper …

3. The 1970 Draft Lottery data sheet (at the end of the exploration) shows the draft number assigned to each of the 366 birthdays.
   a. Find and report the draft number assigned to your birthday.
   b. Is your draft number in the bottom third (1-122), middle third (123-244), or top third (245-366)?

4. The second table at the end of the exploration has ordered the draft numbers within each month.
   a. Use this table to determine the median draft number for your birth month.
   b. Collaborate with your classmates to determine and report the median draft number for all twelve months.

<table>
<thead>
<tr>
<th>Month</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median draft number</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

c. Do you notice any pattern or trend in the median draft numbers over the course of the year? *(Hint: If you do not see a trend, compare the six medians from Jan – June with the 6 medians from July – Dec.)*

5. The correlation coefficient for these data is $r = -0.226$. What does this number reveal? Is it consistent with the scatterplot?

6. a. Suggest two possible explanations (hypotheses) which could have generated the value of the non-zero correlation coefficient.
b. In your own words, how could we go about determining whether random chance is a plausible explanation for the observed correlation value between sequential date and draft number? Explain how the 3S Strategy could be applied here, in particular identifying a simulation strategy you could conduct “by hand.” Note: You do not need to actually carry out a simulation analysis.

i. What is the statistic?

ii. How would you simulate?

iii. How would you evaluate the strength of evidence?

The null hypothesis to be tested is that the lottery was conducted with a fair, random process. The null hypothesis would therefore mean that there is no association between sequential date and draft number for this process. The alternative hypothesis is that this lottery was not conducted with a fair, random process, so there is an association between sequential date and draft number. See FAQ 10.2.1 for further discussion.

How can we assess whether the observed correlation coefficient of $r = -0.226$ is far enough from zero to provide convincing evidence that the lottery process was not random? Like always, we ask how unlikely it would be for a fair, random lottery to produce a correlation value as far from zero as -0.226. Also like always, we answer that question by simulating a large number of fair random lotteries, calculating the correlation coefficient for each one, and seeing how often we obtain a correlation coefficient as or more extreme (as far from zero) as -0.226.

7. Open the Corr/Regression applet. Copy the data from DraftLottery into the applet (remember to include the column titles).

    a. Check the Correlation Coefficient box and confirm that the correlation coefficient is -0.226.

    b. Check the Show Shuffle Options box and select the Correlation radio button to the right to keep track of that statistic. Then press Shuffle Y-values to simulate one fair, random lottery. Record the value of the correlation coefficient between the shuffled draft numbers and sequential date.
c. Press Shuffle Y-values four more times to generate results of four more fair, random lotteries. Record the values of the shuffled correlation coefficients in the table below.

<table>
<thead>
<tr>
<th>Repetition</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation coefficient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

d. Did any of these simulated statistics from fair, random lotteries produce a correlation coefficient as extreme (far from zero) as the observed -0.226?

e. Change the Number of Shuffles from 1 to 995 and press Shuffle Y-values to simulate 995 more fair, random lotteries. Look at the null distribution of these 1000 correlation coefficients. Where is this distribution centered? Why does this make sense?

f. Next to the Count Samples box choose Beyond from the pull-down menu. Specify the observed correlation coefficient (-0.226) and press Count. What proportion, of the 1000 simulated random lotteries produced a correlation coefficient as extreme (as far from zero in either direction) as -0.226? Report the approximate p-value.

g. Interpret this p-value: This is the probability of what, assuming what?

h. What conclusion would you draw from this p-value? Do you have strong evidence that the 1970 draft lottery was not conducted with a fair, random process? Explain the reasoning behind your conclusion.

Once they saw these results, statisticians were quick to point out that something fishy happened with the 1970 draft lottery. The irregularity can be attributed to improper mixing of the balls used in the lottery drawing process. (Balls with birthdays early in the year were placed in the bin first, and balls with birthdays late in the year were placed in the bin last. Without thorough mixing, balls with birthdays late in the year settled near the top of the bin and so tended to be selected earlier.) The mixing process was changed for the 1971 draft lottery (e.g., two bins, one for the draft numbers and one for the birthdays), for which the correlation coefficient turned out to be \( r = 0.014 \).

8. Use your simulation results to approximate the p-value for the 1971 draft lottery. Is there any reason to suspect that this 1971 draft lottery was not conducted with a fair, random process? Explain the reasoning behind your conclusion. Also explain why you don’t need to paste in the data from the 1971 lottery first.
### 1970 Draft Lottery Data

<table>
<thead>
<tr>
<th>date</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>305</td>
<td>86</td>
<td>108</td>
<td>32</td>
<td>330</td>
<td>249</td>
<td>93</td>
<td>111</td>
<td>225</td>
<td>359</td>
<td>19</td>
<td>129</td>
</tr>
<tr>
<td>2</td>
<td>159</td>
<td>144</td>
<td>29</td>
<td>271</td>
<td>298</td>
<td>228</td>
<td>350</td>
<td>45</td>
<td>161</td>
<td>125</td>
<td>34</td>
<td>328</td>
</tr>
<tr>
<td>3</td>
<td>251</td>
<td>297</td>
<td>267</td>
<td>83</td>
<td>40</td>
<td>301</td>
<td>115</td>
<td>261</td>
<td>49</td>
<td>244</td>
<td>348</td>
<td>157</td>
</tr>
<tr>
<td>4</td>
<td>215</td>
<td>210</td>
<td>275</td>
<td>81</td>
<td>276</td>
<td>20</td>
<td>279</td>
<td>145</td>
<td>232</td>
<td>202</td>
<td>266</td>
<td>165</td>
</tr>
<tr>
<td>5</td>
<td>101</td>
<td>214</td>
<td>293</td>
<td>269</td>
<td>364</td>
<td>28</td>
<td>188</td>
<td>54</td>
<td>82</td>
<td>24</td>
<td>310</td>
<td>56</td>
</tr>
<tr>
<td>6</td>
<td>224</td>
<td>347</td>
<td>139</td>
<td>253</td>
<td>155</td>
<td>110</td>
<td>327</td>
<td>114</td>
<td>6</td>
<td>87</td>
<td>76</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>306</td>
<td>91</td>
<td>122</td>
<td>147</td>
<td>35</td>
<td>85</td>
<td>50</td>
<td>168</td>
<td>8</td>
<td>234</td>
<td>51</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>199</td>
<td>181</td>
<td>213</td>
<td>312</td>
<td>321</td>
<td>366</td>
<td>13</td>
<td>48</td>
<td>184</td>
<td>283</td>
<td>97</td>
<td>105</td>
</tr>
<tr>
<td>9</td>
<td>194</td>
<td>338</td>
<td>317</td>
<td>219</td>
<td>197</td>
<td>335</td>
<td>277</td>
<td>106</td>
<td>263</td>
<td>342</td>
<td>80</td>
<td>43</td>
</tr>
<tr>
<td>10</td>
<td>325</td>
<td>216</td>
<td>323</td>
<td>218</td>
<td>65</td>
<td>206</td>
<td>284</td>
<td>21</td>
<td>71</td>
<td>220</td>
<td>282</td>
<td>41</td>
</tr>
<tr>
<td>11</td>
<td>329</td>
<td>150</td>
<td>136</td>
<td>14</td>
<td>37</td>
<td>134</td>
<td>248</td>
<td>324</td>
<td>158</td>
<td>237</td>
<td>46</td>
<td>39</td>
</tr>
<tr>
<td>12</td>
<td>221</td>
<td>68</td>
<td>300</td>
<td>346</td>
<td>133</td>
<td>272</td>
<td>15</td>
<td>142</td>
<td>242</td>
<td>72</td>
<td>66</td>
<td>314</td>
</tr>
<tr>
<td>13</td>
<td>318</td>
<td>152</td>
<td>259</td>
<td>124</td>
<td>295</td>
<td>69</td>
<td>42</td>
<td>307</td>
<td>175</td>
<td>138</td>
<td>126</td>
<td>163</td>
</tr>
<tr>
<td>14</td>
<td>238</td>
<td>4</td>
<td>354</td>
<td>231</td>
<td>178</td>
<td>356</td>
<td>331</td>
<td>198</td>
<td>1</td>
<td>294</td>
<td>127</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td>17</td>
<td>89</td>
<td>169</td>
<td>273</td>
<td>130</td>
<td>180</td>
<td>322</td>
<td>102</td>
<td>113</td>
<td>171</td>
<td>131</td>
<td>320</td>
</tr>
<tr>
<td>16</td>
<td>121</td>
<td>212</td>
<td>166</td>
<td>148</td>
<td>55</td>
<td>274</td>
<td>120</td>
<td>44</td>
<td>207</td>
<td>254</td>
<td>107</td>
<td>96</td>
</tr>
<tr>
<td>17</td>
<td>235</td>
<td>189</td>
<td>33</td>
<td>260</td>
<td>112</td>
<td>73</td>
<td>98</td>
<td>154</td>
<td>255</td>
<td>288</td>
<td>143</td>
<td>304</td>
</tr>
<tr>
<td>18</td>
<td>140</td>
<td>292</td>
<td>332</td>
<td>90</td>
<td>278</td>
<td>341</td>
<td>190</td>
<td>141</td>
<td>246</td>
<td>5</td>
<td>146</td>
<td>128</td>
</tr>
<tr>
<td>19</td>
<td>58</td>
<td>25</td>
<td>200</td>
<td>336</td>
<td>75</td>
<td>104</td>
<td>227</td>
<td>311</td>
<td>177</td>
<td>241</td>
<td>203</td>
<td>240</td>
</tr>
<tr>
<td>20</td>
<td>280</td>
<td>302</td>
<td>239</td>
<td>345</td>
<td>183</td>
<td>360</td>
<td>187</td>
<td>344</td>
<td>63</td>
<td>192</td>
<td>185</td>
<td>135</td>
</tr>
<tr>
<td>21</td>
<td>186</td>
<td>363</td>
<td>334</td>
<td>62</td>
<td>250</td>
<td>60</td>
<td>27</td>
<td>291</td>
<td>204</td>
<td>243</td>
<td>156</td>
<td>70</td>
</tr>
<tr>
<td>22</td>
<td>337</td>
<td>290</td>
<td>265</td>
<td>316</td>
<td>326</td>
<td>247</td>
<td>153</td>
<td>339</td>
<td>160</td>
<td>117</td>
<td>9</td>
<td>53</td>
</tr>
<tr>
<td>23</td>
<td>118</td>
<td>57</td>
<td>256</td>
<td>252</td>
<td>319</td>
<td>109</td>
<td>172</td>
<td>116</td>
<td>119</td>
<td>201</td>
<td>182</td>
<td>162</td>
</tr>
<tr>
<td>24</td>
<td>59</td>
<td>236</td>
<td>258</td>
<td>2</td>
<td>31</td>
<td>358</td>
<td>23</td>
<td>36</td>
<td>195</td>
<td>196</td>
<td>230</td>
<td>95</td>
</tr>
<tr>
<td>25</td>
<td>52</td>
<td>179</td>
<td>343</td>
<td>351</td>
<td>361</td>
<td>137</td>
<td>67</td>
<td>286</td>
<td>149</td>
<td>176</td>
<td>132</td>
<td>84</td>
</tr>
<tr>
<td>26</td>
<td>92</td>
<td>365</td>
<td>170</td>
<td>340</td>
<td>357</td>
<td>22</td>
<td>303</td>
<td>245</td>
<td>18</td>
<td>7</td>
<td>309</td>
<td>173</td>
</tr>
<tr>
<td>27</td>
<td>355</td>
<td>205</td>
<td>268</td>
<td>74</td>
<td>296</td>
<td>64</td>
<td>289</td>
<td>352</td>
<td>233</td>
<td>264</td>
<td>47</td>
<td>78</td>
</tr>
<tr>
<td>28</td>
<td>77</td>
<td>299</td>
<td>223</td>
<td>262</td>
<td>308</td>
<td>222</td>
<td>88</td>
<td>167</td>
<td>257</td>
<td>94</td>
<td>281</td>
<td>123</td>
</tr>
<tr>
<td>29</td>
<td>349</td>
<td>285</td>
<td>362</td>
<td>191</td>
<td>226</td>
<td>353</td>
<td>270</td>
<td>61</td>
<td>151</td>
<td>229</td>
<td>99</td>
<td>16</td>
</tr>
<tr>
<td>30</td>
<td>164</td>
<td>217</td>
<td>208</td>
<td>103</td>
<td>209</td>
<td>287</td>
<td>333</td>
<td>315</td>
<td>38</td>
<td>174</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>211</td>
<td>30</td>
<td>313</td>
<td>193</td>
<td>11</td>
<td>79</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The following table arranges the draft numbers in order for each month:

<table>
<thead>
<tr>
<th>rank</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>4</td>
<td>29</td>
<td>2</td>
<td>31</td>
<td>20</td>
<td>13</td>
<td>11</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>52</td>
<td>25</td>
<td>30</td>
<td>14</td>
<td>35</td>
<td>22</td>
<td>15</td>
<td>21</td>
<td>6</td>
<td>7</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>58</td>
<td>57</td>
<td>33</td>
<td>32</td>
<td>37</td>
<td>28</td>
<td>23</td>
<td>36</td>
<td>8</td>
<td>24</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>68</td>
<td>108</td>
<td>62</td>
<td>40</td>
<td>60</td>
<td>27</td>
<td>44</td>
<td>18</td>
<td>38</td>
<td>46</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>77</td>
<td>86</td>
<td>122</td>
<td>74</td>
<td>55</td>
<td>64</td>
<td>42</td>
<td>45</td>
<td>49</td>
<td>72</td>
<td>47</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>92</td>
<td>89</td>
<td>136</td>
<td>81</td>
<td>65</td>
<td>69</td>
<td>50</td>
<td>48</td>
<td>63</td>
<td>79</td>
<td>51</td>
<td>39</td>
</tr>
<tr>
<td>7</td>
<td>101</td>
<td>91</td>
<td>139</td>
<td>83</td>
<td>75</td>
<td>73</td>
<td>67</td>
<td>54</td>
<td>71</td>
<td>87</td>
<td>66</td>
<td>41</td>
</tr>
<tr>
<td>8</td>
<td>118</td>
<td>144</td>
<td>166</td>
<td>90</td>
<td>103</td>
<td>85</td>
<td>88</td>
<td>61</td>
<td>82</td>
<td>94</td>
<td>76</td>
<td>43</td>
</tr>
<tr>
<td>9</td>
<td>121</td>
<td>150</td>
<td>169</td>
<td>124</td>
<td>112</td>
<td>104</td>
<td>93</td>
<td>102</td>
<td>113</td>
<td>117</td>
<td>80</td>
<td>53</td>
</tr>
<tr>
<td>10</td>
<td>140</td>
<td>152</td>
<td>170</td>
<td>147</td>
<td>130</td>
<td>109</td>
<td>98</td>
<td>106</td>
<td>119</td>
<td>125</td>
<td>97</td>
<td>56</td>
</tr>
<tr>
<td>11</td>
<td>159</td>
<td>179</td>
<td>200</td>
<td>148</td>
<td>133</td>
<td>110</td>
<td>115</td>
<td>111</td>
<td>149</td>
<td>138</td>
<td>99</td>
<td>70</td>
</tr>
<tr>
<td>12</td>
<td>164</td>
<td>181</td>
<td>213</td>
<td>191</td>
<td>155</td>
<td>134</td>
<td>120</td>
<td>114</td>
<td>151</td>
<td>171</td>
<td>107</td>
<td>78</td>
</tr>
<tr>
<td>13</td>
<td>186</td>
<td>189</td>
<td>217</td>
<td>208</td>
<td>178</td>
<td>137</td>
<td>153</td>
<td>116</td>
<td>158</td>
<td>176</td>
<td>126</td>
<td>84</td>
</tr>
<tr>
<td>14</td>
<td>194</td>
<td>205</td>
<td>223</td>
<td>218</td>
<td>183</td>
<td>180</td>
<td>172</td>
<td>141</td>
<td>160</td>
<td>192</td>
<td>127</td>
<td>95</td>
</tr>
<tr>
<td>15</td>
<td>199</td>
<td>210</td>
<td>239</td>
<td>219</td>
<td>197</td>
<td>206</td>
<td>187</td>
<td>142</td>
<td>161</td>
<td>196</td>
<td>131</td>
<td>96</td>
</tr>
<tr>
<td>16</td>
<td>211</td>
<td>212</td>
<td>256</td>
<td>231</td>
<td>226</td>
<td>209</td>
<td>188</td>
<td>145</td>
<td>175</td>
<td>201</td>
<td>132</td>
<td>100</td>
</tr>
<tr>
<td>17</td>
<td>215</td>
<td>214</td>
<td>258</td>
<td>252</td>
<td>250</td>
<td>222</td>
<td>190</td>
<td>154</td>
<td>177</td>
<td>202</td>
<td>143</td>
<td>105</td>
</tr>
<tr>
<td>18</td>
<td>221</td>
<td>216</td>
<td>259</td>
<td>253</td>
<td>276</td>
<td>228</td>
<td>193</td>
<td>167</td>
<td>184</td>
<td>220</td>
<td>146</td>
<td>123</td>
</tr>
<tr>
<td>19</td>
<td>224</td>
<td>236</td>
<td>265</td>
<td>260</td>
<td>278</td>
<td>247</td>
<td>227</td>
<td>168</td>
<td>195</td>
<td>229</td>
<td>156</td>
<td>128</td>
</tr>
<tr>
<td>20</td>
<td>235</td>
<td>285</td>
<td>267</td>
<td>262</td>
<td>295</td>
<td>249</td>
<td>248</td>
<td>198</td>
<td>204</td>
<td>234</td>
<td>174</td>
<td>129</td>
</tr>
<tr>
<td>21</td>
<td>238</td>
<td>260</td>
<td>269</td>
<td>296</td>
<td>272</td>
<td>270</td>
<td>245</td>
<td>207</td>
<td>237</td>
<td>182</td>
<td>135</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>251</td>
<td>292</td>
<td>275</td>
<td>271</td>
<td>298</td>
<td>274</td>
<td>277</td>
<td>261</td>
<td>225</td>
<td>241</td>
<td>185</td>
<td>157</td>
</tr>
<tr>
<td>23</td>
<td>280</td>
<td>297</td>
<td>293</td>
<td>273</td>
<td>308</td>
<td>301</td>
<td>279</td>
<td>286</td>
<td>232</td>
<td>243</td>
<td>203</td>
<td>162</td>
</tr>
<tr>
<td>24</td>
<td>305</td>
<td>299</td>
<td>300</td>
<td>312</td>
<td>313</td>
<td>335</td>
<td>284</td>
<td>291</td>
<td>233</td>
<td>244</td>
<td>230</td>
<td>163</td>
</tr>
<tr>
<td>25</td>
<td>306</td>
<td>302</td>
<td>317</td>
<td>316</td>
<td>319</td>
<td>341</td>
<td>287</td>
<td>307</td>
<td>242</td>
<td>254</td>
<td>266</td>
<td>165</td>
</tr>
<tr>
<td>26</td>
<td>318</td>
<td>338</td>
<td>323</td>
<td>336</td>
<td>321</td>
<td>353</td>
<td>289</td>
<td>311</td>
<td>246</td>
<td>264</td>
<td>281</td>
<td>173</td>
</tr>
<tr>
<td>27</td>
<td>325</td>
<td>347</td>
<td>332</td>
<td>340</td>
<td>326</td>
<td>356</td>
<td>303</td>
<td>324</td>
<td>255</td>
<td>283</td>
<td>282</td>
<td>240</td>
</tr>
<tr>
<td>28</td>
<td>329</td>
<td>363</td>
<td>334</td>
<td>345</td>
<td>330</td>
<td>358</td>
<td>322</td>
<td>333</td>
<td>257</td>
<td>288</td>
<td>309</td>
<td>304</td>
</tr>
<tr>
<td>29</td>
<td>337</td>
<td>365</td>
<td>343</td>
<td>346</td>
<td>357</td>
<td>360</td>
<td>327</td>
<td>339</td>
<td>263</td>
<td>294</td>
<td>310</td>
<td>314</td>
</tr>
<tr>
<td>30</td>
<td>349</td>
<td>354</td>
<td>351</td>
<td>361</td>
<td>366</td>
<td>331</td>
<td>344</td>
<td>315</td>
<td>342</td>
<td>348</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>355</td>
<td>362</td>
<td>364</td>
<td>350</td>
<td>352</td>
<td>359</td>
<td>328</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Section 10.2 Summary

The 3S Strategy can be used to assess whether a sample correlation coefficient is extreme enough to provide strong evidence that the variables are associated in the population.

- The null hypothesis is that there is no association between the two variables in the population.
- The statistic used to measure the (linear) association is the correlation coefficient.
- Shuffling the values of the response variable is used to produce simulated values of the statistic.
  - This shuffling simulates values of the statistic under the assumption of no underlying association between the two variables.
- As always, the p-value is calculated as the proportion of repetitions in which the simulated value of the statistic is at least as extreme as the observed value of the statistic.
  - Also, as always, a small p-value provides evidence against the null hypothesis in favor of the alternative hypothesis.
  - We can test for positive or negative linear associations (one-sided p-values) or for any association (two-sided p-value).

Section 10.3: Least Squares Regression

In the last two sections we used graphical and numerical methods to describe the observed association between two quantitative variables, and then looked at a simulation method for drawing inferences based on the correlation coefficient. If we decide the association is linear, it is helpful to develop a mathematical model of that association. This is especially helpful in allowing us to make predictions about the response variable from the explanatory variable. The most common way to do this is with a least squares regression line (often called the best fit line, regression line or, simply, regression) which is the line that gets as “close” as possible to all of the data points.

Example 10.3: Are Dinner Plates Getting Larger (revisited)?

In Exploration 10.1 we looked at a recent study investigating a potential association between the year of manufacture and dinner plate size, finding strong evidence of a positive association, meaning that, based on a random sample of plates for sale on eBay on a given day, plate sizes appear to be increasing over time. A scatterplot of the data is presented again in Figure 10.7.
This scatterplot reveals a moderately strong positive association, as there does appear to be a trend of increasing plate sizes in more recent years. This observation is supported by the correlation coefficient ($r = 0.604$). The association also appears to be roughly linear, indicating that a line would do a reasonable job of summarizing the overall relationship. We have added such a line, specifically the least squares regression line, to the scatterplot in Figure 10.7.

**Figure 10.7:** A scatterplot and least squares regression line for the size of American dinner plates and the year of manufacture

![Scatterplot and Regression Line](image)

As you may recall, lines can be summarized by an equation which indicates the slope and $y$-intercept of the line. The line shown above has a slope of $0.0128$ and a $y$-intercept of $-14.80$.

**Notation:** The equation of the best fit line is written as $\hat{y} = a + b(x)$, where
- $a$ is the $y$-intercept and $b$ is the slope
- $x$ is a value of the explanatory variable
- $\hat{y}$ is the predicted value for the response variable

The least squares line for the dinner plate data turns out to be $\hat{y} = -14.8 + 0.0128x$, or better written as $\hat{\text{diameter}} = -14.8 + 0.0128(\text{year})$.

Such an equation allows us to predict the plate diameter in a particular year. If we substitute 2000 in for the year, we predict a plate diameter of $-14.8 + 0.0128(2000) = 10.8$ inches.

**Think about it:** What do you predict for a plate diameter for a plate manufactured in the year 2001? How
The slope of this equation is $b = 0.0128$. This means that, according to our line, dinner plate diameters are predicted to increase by 0.0128 inches per year on average. Your prediction for 2001 should have been roughly 0.0128 inches larger than your prediction for 2000. This is true no matter what two years you use, as long as they are one year apart. If you wanted to predict the change in average plate size associated with a 10 year difference in manufacturing dates, you could simply multiply the slope by 10.

**Key idea:** The slope of a least squares regression line is interpreted as the predicted change in the average response variable for a one-unit change in the explanatory variable.

Note that the slope of the regression line for the dinner plate data is positive, as was the correlation coefficient. These two statistics will always have the same sign.

**Key idea:** For a given dataset, the signs (positive, negative, or zero) for the correlation coefficient and the slope of the regression line must be the same.

You can think of the $y$-intercept in a couple of different ways. Graphically, it is where the regression line crosses the $y$-axis. Because the $x$-value of the $y$-axis is 0, the $y$-intercept can also be thought of as the $\hat{y}$-value (or predicted response) when the $x$-value (or explanatory variable) equals 0. We had a $y$-intercept of -14.8 in the dinner plate equation. This tells us that, according to the line, the predicted size of American dinner plates in year 0 was -14.8 inches. Now this doesn’t make much sense, does it? For one, we can’t have a negative value for the size of dinner plates. Our equation might work reasonably well within the range of values of the explanatory variable with our given data, but may fail miserably outside that range. In fact, our equation should only be used to predict the size of dinner plates from about 1950 to 2010. We don’t know whether the same relationship between year and plate size holds outside this range, so we should not make predictions outside this range. Going way back to year 0, for example, gives us an answer that makes no sense at all.

**Key idea:** Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is known as *extrapolation*, and can give very misleading predictions.

We should also make sure the interpretation is meaningful in the study context. In many situations the intercept can have a meaningful interpretation, such as the start-up (time = 0) cost of a manufacturing process.
**Key idea:** The $y$-intercept of a regression line is interpreted as the predicted value of the response variable when the explanatory variable has a value of zero (though be wary of extrapolation in interpreting the intercept or other values outside the original data range).

**What makes this line the “best”?**

We’ve talked about how the line above is the “best” line. But what makes it the best? Intuitively, we said that the best fitting line gets as close as possible to all of the points, but we can be more precise by better understanding the idea of a residual. The vertical distance from a point to the line is called a residual. So the residuals measure the difference between the predicted and observed value of the response variable (in this case, plate size) for each $x$ value. Figure 10.8 shows the residuals for the plate example.

**Key idea:** A *residual* is the difference between an observed response and the corresponding prediction made by the least squares regression line ($\text{residual} = \text{observed} - \text{predicted}$). Thus, negative residuals occur when points are below the best fit line and positive residuals occur when points are above the best fit line.

**Figure 10.8:** Residuals for best-fit line of year vs. plate size

Each observational unit (plate) has a residual which can be found by subtracting the predicted plate size from the observed (real) plate size (in other words finding the value of $y - \hat{y}$). For example, the plate manufactured in 1950 has a predicted plate size of 10.16 inches, but is actually 10 inches, and so the residual is $10 - 10.16 = -0.16$ inches.
There are actually a number of ways to get “as close” to all of the points as possible, one way would be to add up all of the values of the residuals (making them positive first, so they didn’t cancel out), and finding the line that minimized the sum of the (absolute) residuals. In practice, however, finding such a line is very difficult. Instead, we will find the line that minimizes the sum of the squared residuals.

**Key Idea:** The least squares regression line minimizes the sum of squared residuals.

Although the reasons for this change are due to calculus (and not something we really want you to worry about!), the implications aren’t too different than if we were minimizing the sum of the absolute residuals. The only subtle difference is that because we are squaring, unusual observations (extreme in the $x$ direction) have more “pull” as the squared value of the residual is larger relative to more typical points. Thus, these kinds of points can be particularly influential — just like they were when we explored the correlation coefficient. See FAQ 10.3.1 for more.

**Key idea:** An observation or set of observations is considered influential if removing the observation from the dataset substantially changes the values of the correlation coefficient and/or the least squares regression equation. Typically, observations that have extreme explanatory variable values (far below or far above $x$) are potentially influential. They may not have large residuals, having pulled the line close to them.

One final statistic that is of interest in a linear regression is called the **coefficient of determination** ($r^2$). This statistic measures the percentage of total variation in the response variable that is explained by the linear relationship with the explanatory variable. The value is literally the square of the correlation coefficient. In this example, the correlation coefficient is 0.604, and so the coefficient of determination is $(0.604)^2 \approx 0.365$, meaning that 36.5% of the observed variation in plate sizes is explained by knowing the year of manufacture.

Another way to compute the coefficient of determination is as follows:

$$r^2 = \frac{\text{SSE} (\bar{y}) - \text{SSE} (\text{regression line})}{\text{SSE} (\bar{y})}$$

where $\text{SSE}(\bar{y})$ is the sum of the squared residuals for the least squares regression line and $\text{SSE}(\bar{y})$ is the sum of squared residuals for the horizontal line at the average value of the response variable. We can think of $\text{SSE}(\bar{y})$ as a measure of the overall variability in the response, and we are seeing how much less variability there is about the regression line compared to this “no association” line.

There is a third way to think about the coefficient of determination. If we look at the plate sizes alone, variability of the plate sizes, as measured by the standard deviation is 0.430. We can also use the squared
standard deviation or variance to reflect this, $0.430^2 \approx 0.185$. But if we look at the residuals, the “unexplained variability” is down to 0.343. That is, the standard deviation of the residuals is 0.343, meaning that the variance of the residuals is $0.343^2 \approx 0.118$, a reduction of $0.185 – 0.118 = 0.067$. So we’ve explained about $0.067/0.430$ or 36% of the original variability in the plate sizes.

**Figure 10.9a. Variability in plate sizes**

<table>
<thead>
<tr>
<th>size</th>
<th>10</th>
<th>10.5</th>
<th>11</th>
<th>11.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(n=20)
SD = 0.430

**Figure 10.9b. Variability in residuals**

<table>
<thead>
<tr>
<th>residuals</th>
<th>-0.8</th>
<th>-0.4</th>
<th>0</th>
<th>0.4</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(n=20)
SD = 0.343

**Key idea:** The coefficient of determination ($r^2$) is the percentage of total observed variation in the response variable that is accounted for by changes (variability) in the explanatory variable.

**FAQ 10.3.1 Why least squares?**

Q: Why least squares? Why do we square the deviations? Why not absolute values?

A: Not only is your question a thoughtful one, it’s also a question that invites us back in history. In the 1700s, scientists tried to do what you suggest – to minimize the sum of absolute deviations: they wanted to use least MAD(deviation) instead of least sum of squared deviations. The scientific logic was good, but the math turned out to be too messy. Sometimes it worked, and there was a single “best” line that minimized the MAD statistic. But sometimes there would be infinitely many “best” lines, with no way to choose from among them. “Best” and “infinitely many” don’t go together. Even when there was a single best line, it was devilishly hard to find. Overall, the math was just too hard.

Q: So least squares is pretty much just the fallback Plan B after least absolute deviations failed?

A: Yes. The justification for least absolute deviations makes good sense, but the theory is much simpler for least squares. As Fred Mosteller, one of the greatest statisticians of the last century, used to tell his students at Harvard, “Mathematicians love calculus, and calculus loves squares.”
Q: Does the choice – absolute values or squares -- make much difference in the line you get?

A: Another good question. Sometimes not, but sometimes it does.

Q: Say more.

A: Remember SD and IQR and outliers?

Q: You mean SD is sensitive to outliers and IQR is resistant?

A: Exactly. Remember why?

Q: Because the SD squares the deviations, and the IQR does not. Squaring gives extra weight to large deviations.

A: Apply the same logic to the least squares line.

Q: Outliers have a big effect because their large deviations get extra weight from the squaring.

A: Bravo! This is why some observations can become influential points – the least squares line won’t let their residuals get too large. If this were second grade in the 1950s you’d get to stay after class and dust the erasers. If this were the 1970s and you were a statistician, you’d be able to get a big research grant from the National Science Foundation. If this was Dr. Bastian’s lab and you were a dolphin, you’d have just earned a fish.

**Exploration 10.3: Predicting Height from Footprints**

Can a footprint taken at the scene of a crime help to predict the height of the criminal? In other words, is there an association between height and foot length? To investigate this question, a sample of 20 statistics students measured their height (in inches) and their foot length (in centimeters).

1. Let's think about this study.
   a. What are the observational units in this study?
b. Identify the two variables recorded for each observational unit. Which is the explanatory variable, and which is the response? Also classify these variables as categorical or quantitative.

Explanatory: Type:

Response: Type:

c. Is this an observational study or a randomized experiment?

The sample data for the 20 students appear in the Corr/Regression applet.

2. Open the applet, and look at the scatterplot for the data already there. Describe the association between the variables as revealed in the scatterplot. (Hint: Remember to comment on direction, strength, and form of association as well as unusual observations.)

3. Would you say that a straight line could summarize the relationship between height and foot length reasonably well?

4. Check the Show Movable Line box to add a blue line to the scatterplot. If you place your mouse over one of the green squares at the ends of the line and drag, you can change the slope of the line and move it. You can also use the mouse to move the green dot up and down vertically to change the intercept of the line.

   a. Move the line until you believe your line “best” summarizes the relationship between height and foot length for these data. Write down the resulting equation for your line (using traditional statistical notation).

   b. Why do you believe that your line is “best?”
c. Did all students in your class obtain the same line/equation?

d. How can we decide whether your line provides a better fit to the data than other students’ lines? Suggest a criterion for deciding which line “best” summarizes the relationship.

One way to draw the best fit line is to minimize the distance of the points to the line (these distances are called residuals).

**Key idea:** A *residual* is the difference between an observed response and the corresponding prediction made by the least squares regression line \((\text{residual} = \text{observed} - \text{predicted})\). Thus, negative residuals occur when points are below the best fit line and positive residuals occur when points are above the best fit line.

5. Would points above the line have a positive or negative residual, or is it impossible to tell? What about points below the line?

   Above line:    Below line:

6. Check the **Show Residuals** box to visually represent these residuals for your line on the scatterplot. The applet also reports the sum of the values of the residuals (SAE). SAE stands for “Sum of the Absolute Errors.” The acronym indicates that we need to make residuals positive before we add them up and that sometimes people call residuals “errors.”

   Record the SAE value for your line ___________

   What is the best (lowest) SAE in the class?_________________

7. It turns out that a more common criterion for determining the “best” line is to instead look at the sum of the squared residuals (SSE). This approach is similar to simply adding up the residuals, but is even more strict in not letting individual residuals get too large.

   Check the **Show Squared Residuals** box to visually represent the squared residual for each observation. Note that we can visually represent the squared residual as the area of a square where each side of the square has length equal to the residual.

   a. What is the SSE (sum of squared residuals) for your line?____________

   What is the best (lowest) SSE in the class?_________
b. Now continue to adjust your line until you think you have minimized the sum of the squared residuals.

Report your new equation_________________________________

Report your new SSE value_____________________________

What is the best SSE in the class?_________________________

**Key Idea:** The least squares regression line minimizes the sum of squared residuals.

8. Now check the **Show Regression Line** box to determine and display the equation for the line that actually does minimize (as can be shown using calculus) the sum of the squared residuals.

   a. Record the equation of the least squares regression line by indicating the appropriate slope and intercept of the line. Note that we’ve used variable names in the equation, not generic \( x \) and \( y \). And put a carat (“hat”) over the \( y \) variable name to emphasize that the line gives predicted values of the \( y \) (response) variable.

   \[ \hat{\text{Height}} = \square \text{FootLength} \]

   **Notation:** The equation of the best fit line is written as \( \hat{y} = a + b(x) \) where
   - \( a \) is the y-intercept
   - \( b \) is the slope
   - \( x \) is a value of the explanatory variable
   - \( \hat{y} \) is the predicted value for the response variable

   b. Did everyone in your class obtain the same equation?

   c. Is the slope positive or negative? Explain how the sign of the slope tells you about whether your data display a positive or a negative association.
Key idea: For a given dataset, the signs (positive or negative) for the correlation coefficient and the slope of the regression line must be the same.

9.
   a. Use the least squares regression line to predict the height of someone whose foot length is 28 cm. (Simply plug in the value of 28 cm for foot length in the equation of the line.)

   b. Use the least squares regression line to predict the height of someone whose foot length is 29 cm.
   c. By how much do your predictions in (a) and (b) differ? Does this number look familiar? Explain.

Key idea: The slope coefficient of a least squares regression model is interpreted as the predicted change in the response (y) variable for a one-unit change in the explanatory (x) variable.

d. Interpret the slope in context:

   The slope of the regression line predicting height based on foot length is ________, meaning that for every additional _______ cm increase in foot length, their predicted height increases by ________ inches.

10.
   a. Use the least squares regression line to predict the height of someone who has a foot length of 0 cm.

   b. Your answer to (a) should look familiar. What is this value?
**Key idea:** The y-intercept of a regression line is interpreted as the predicted value of the response variable when the explanatory variable has a value of zero (though be wary of extrapolation in interpreting the intercept or other values outside the original data range).

**Key idea:** Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is known as *extrapolation* and can give very misleading predictions.

c. Explain how your prediction of the height of someone whose foot length is 0 cm is an example of extrapolation.

11. Earlier, we explored the notion of a residual as the distance of a point to a line.

   a. Using the equation for the best fit line that you reported in #8, find the residual for the person whose foot length is 32 centimeters and height is 74 inches. Note: Find this value by taking the actual height and subtracting the person’s predicted height from the equation.

   b. Is this person’s dot on the scatterplot above or below the line? How does the residual tell you this?

Residuals are helpful for identifying unusual observations. But not all observations of interest have large residuals.

12. Uncheck the **Show Movable Line** box to remove it from the display and check the **Move observations** box.
a. Now click on one student’s point in the scatterplot and drag the point up and down (The original regression line will remain in grey.) Does the regression line change much as you change this student’s height?

b. Repeat the previous question, using a student with a very small $x$ (foot size) value and then a point with an $x$ value near the middle and then a point with a very large $x$ value. Which of these seem(s) to have the most influence on the least squares regression line? Explain.

Key idea: An observation or set of observations is considered influential if removing the observation from the dataset substantially changes the values of the correlation coefficient and/or the least squares regression equation. Typically, observations that have extreme explanatory variable values (far below or far above $x$) are potentially influential. They may not have large residuals, having pulled the line close to them.

Residuals also help us measure how accurate our predictions are from using the regression line. In particular, we can compare the “prediction errors” from the regression line to the prediction errors if we made no use of the explanatory variable.

13. Press the Revert button to reload the original data set. Recheck the Show Movable Line box to redisplay the blue line. Notice that this line is flat at the mean of the $y$ (height) values.

a. Check the Show Squared Residuals box (under the Movable Line information) to determine the SSE if we were to use the average height ($\bar{y}$) as the predicted value for every $x$ (foot size). Record this value.

b. What is the slope of this line?

c. If the slope of the best fit line is zero, our data shows ________________ (positive/negative/no) association between the explanatory and response variables.
14. Now compare this to the SSE value from the regression line. Which is smaller? Why does that make sense?

**Coefficient of Determination \( (r^2) \)**

15. A quantity related to the correlation coefficient is called the coefficient of determination, or \( r^2 \).

a. To measure how much the regression line decreases the “unexplained variability” in the response variable, we can calculate the percentage reduction in SSE.

\[
x^2 = 100\% \times \frac{\text{SSE}(\bar{y}) - \text{SSE (regression line)}}{\text{SSE} (\bar{y})}
\]

**Key Idea:** The coefficient of determination indicates the percentage observed in the response variable (height) that is explained by changes in the explanatory variable (foot length).

b. Complete the following statement: The coefficient of determination is __________ percent, this means that __________ percent of the variation in people’s ________________ is attributable to changes in their ______________________________.

If we think of \( \hat{y} = \bar{y} \) as the “worst” line (no association), then we are seeing how much better the actual line is than this worst line.

c. Find the value of the correlation coefficient using the applet, and confirm that when you square the correlation coefficient you get the same number as the applet reports for the coefficient of determination.

For more on the coefficient of determination see Example 10.3.
Section 10.3 Summary

When the association between two quantitative variables has a linear form, a line can be used as a mathematical model for summarizing the relationship and for predicting values of the response variable from the explanatory variable.

- The least squares regression line selects the intercept and slope coefficients to minimize the sum of the squared vertical deviations between the observations (y values) and the line.
  - These vertical deviations between the observed y values and the line are called residuals.
  - A primary use of the least squares regression line is to make predictions for the response variable based on the explanatory variable.
    - Making predictions for values of the explanatory variable not included in the range of our data values is called extrapolation and is often unwise.
  - The slope coefficient is interpreted as the predicted change in the response variable associated with a one-unit change in the explanatory variable.
  - The intercept coefficient is interpreted as the predicted value of the response variable when the explanatory variable equals zero.
    - In many contexts the interpretation of the intercept is not sensible and/or may be extrapolation.
  - The square of the correlation coefficient ($r^2$) indicates the proportion of variability in the response variable that is explained by the least squares line with the explanatory variable.
    - This is sometimes called the coefficient of determination.

- The sign of the slope (positive, negative, or zero) matches the sign of the correlation coefficient (positive, negative, or zero).

- Influential observations affect the regression line substantially.
  - Influential observations typically have extreme values of the explanatory variable.
  - Outliers in a regression setting have large residuals.
  - Influential observations are typically not outliers because they pull the line toward themselves.

Section 10.4: Inference for Regression Slope: Simulation-Based Approach
In the previous section, we looked at how least squares linear regression can be used to describe a linear relationship between two quantitative variables. In that section, we mentioned that the sign on the slope of the regression equation and the sign on the correlation coefficient are always the same. For example, when there is a positive association in the data, both the correlation coefficient and the slope of the regression equation are positive. In Section 10.2, we saw how to use the sample correlation coefficient in a simulated-based test about a null hypothesis of no association. In this section, we will see how we can do the same type of inference but now with the population slope as the parameter of interest.

**Example 10.4: Do students who spend more time in non-academic activities, tend to have lower GPAs?**

College students have lots of opportunities to be involved in different activities. Some also hold jobs or participate in athletics. Both of these can be very time consuming. Time just socializing with friends also takes up lots of time for some. Does time spent on these types of activities have a negative association with students’ GPAs? Many researchers have conducted studies to answer this question. We will look at results from a study carried out by a couple of undergraduate students at the University of Minnesota (Ock, 2008). They surveyed 42 students (34 from a research methods class and 8 from a business management class). They asked questions about time spent per week on various nonacademic activities like work, watching TV, exercising, and socializing. They calculated the total time spent on all nonacademic activities and compared that with the students’ GPAs. Those data, gleaned from a scatterplot in their paper, are presented in Figure 10.10, along with the least squares regression line.
Figure 10.10: A scatterplot of students’ GPAs and the total number of hours spent on nonacademic activities for 42 students

Think about it: Describe the nature of the association between GPA and hours spent on nonacademic activities revealed by this scatterplot.

We discussed outliers briefly earlier. In the regression setting, an outlier is defined as an observation with a large residual. It is often useful to identify the outliers separately, if possible, and then describe the resulting association as if the outliers were removed from the dataset. Apart from the two low outliers (two identified students with GPAs below 2.0), these data exhibit a moderate to weak negative association. The correlation coefficient for the data in Figure 10.9 is \( r = -0.290 \) and the regression equation is:

\[
\text{GPA} = 3.60 - 0.0059(\text{nonacademic hours}).
\]

As you can see, for the sample of 42 students both the correlation coefficient and the slope of the regression equation are negative.

Think about it: Interpret the slope and intercept of this regression line in the context of non-academic hours and GPA.

The slope of -0.0057 indicates that the predicted GPA decreases by 0.0057 grade points for each additional hour spent on nonacademic activities per week. The intercept reveals that the predicted GPA for a student who spends absolutely no time on nonacademic activities is 3.60.
What can we infer about the larger population represented by this sample? Are we convinced that the slope in the population is negative, as we suspected in advance? Or is it plausible that the GPA values we observed are just arbitrarily paired up with the hours spent on nonacademic activities, and that the association we observed in this sample arose just by chance? Our hypotheses are as follows.

Null Hypothesis: There is no association between the number of hours students spend on nonacademic activities and student GPA in the population.

Alternative Hypothesis: There is a negative association between the number of hours students spend on nonacademic activities and student GPA in the population.

In Section 10.2 we tested the statistical significance of an association between two quantitative variables using the correlation coefficient as our statistic. In this example we will use the slope of the sample regression equation as our statistic. We previously said that a regression line can be represented by the equation $\hat{y} = a + b(x)$ where $a$ is the $y$-intercept and $b$ is the slope. This is the form of the equation that represents a sample. For this reason, both the $y$-intercept ($a$) and slope ($b$) are statistics.

We will now use the sample slope as the statistic in test of no association. To do this we will use the Corr/Regression applet we used for correlation coefficients. We do this exactly the same way as was done with correlation except select the slope radio button so that slope is used as our statistic instead of the correlation coefficient.

In particular, we shuffle the values of the response variable in the data table, which is equivalent to “dealing” the response variable values out randomly to the values of the explanatory variable in the data table. We did 1000 shuffles and the results are shown in Figure 10.11. See FAQ 10.4.1. for more on the unique behavior of the 1000 shuffled lines. We look for the lower-tail p-value to coincide with our one sided alternative hypothesis.
**Figure 10.11:** Assuming there was no genuine association between number of hours of nonacademic activities and GPA, we did 1000 shuffles of the response variable and calculated the slope of the resulting regression equation each time.

Here is a summary of our simulation model:

**Null hypothesis** = No association between x and y variables

**One repetition** = Re-randomizing the response outcomes to the explanatory variable outcomes

**Statistic** = slope coefficient

From our 1000 shuffles we found 33 simulated slopes that were as small or smaller as our observed sample slope of -0.0057. Thus our simulated p-value is 0.033, and we can conclude there is strong evidence of a negative association between the number of hours students spend on nonacademic activities and their GPA for this population. If we had used the correlation coefficient for our statistic instead of the slope, our results would have been exactly the same (unlike what we saw in Chapters 8 and 9 when we used the MAD statistic as an alternative to a chi-squared or ANOVA F-statistic). We will revisit this idea in the next section.

**Key idea:** For a given dataset, a test of association based on a slope is equivalent to a test of association based on a correlation coefficient.

We must ask to what population we can make this inference. The researchers used a convenience sample of mostly Research Methods students. They also added a few business management students in order for the sample to more closely match the racial demographics of the undergraduates at the University of
Minnesota. It would seem they did this because their population of interest was all undergraduates at the university. Do you think the sample collected is representative of all students at the university in terms of the association that was being studied? If we had a simple random sample of all students, then we could easily make this connection. A convenience sample, however, brings with it some concerns. When we are talking about student behavior, it is probably not safe to assume that a convenience sample will produce representative results. We should be very cautious as to exactly what population we can make this inference. Perhaps all social science students (the type that would take a research methods class) might be okay, but beyond that we would have a pretty weak argument.

Finally, we should note that because this study is observational, we cannot conclude that time spent on nonacademic activities causes a decrease in GPA.

**Exploration 10.4: Perceptions of Heaviness**

Researchers believe a person’s body is used as a perceptual ruler and people will judge the size of an object based on its relationship to parts of their body. Specifically, some researchers thought people with smaller hands will perceive objects to be bigger and hence heavier than those with larger hands. Linkenauger, Mohler, and Proffitt (2011) collected data on 46 participants, recording their hand width and estimated weight of bean bags. The results are shown in Table 10.4.

**Table 10.4: Hand widths (in centimeters) and estimated weight of bean bags (in grams) for 46 participants**

<table>
<thead>
<tr>
<th>Hand Width</th>
<th>Estimated Weight</th>
<th>Hand Width</th>
<th>Estimated Weight</th>
<th>Hand Width</th>
<th>Estimated Weight</th>
<th>Hand Width</th>
<th>Estimated Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.4</td>
<td>75.2</td>
<td>7.7</td>
<td>89.2</td>
<td>8.6</td>
<td>96.8</td>
<td>9.2</td>
<td>92.1</td>
</tr>
<tr>
<td>7.4</td>
<td>83.6</td>
<td>7.7</td>
<td>83.6</td>
<td>8.6</td>
<td>103.1</td>
<td>9.2</td>
<td>95.9</td>
</tr>
<tr>
<td>7.4</td>
<td>86.2</td>
<td>8.0</td>
<td>97.0</td>
<td>8.6</td>
<td>111.9</td>
<td>9.2</td>
<td>103.0</td>
</tr>
<tr>
<td>7.4</td>
<td>92.6</td>
<td>8.0</td>
<td>97.0</td>
<td>8.6</td>
<td>112.5</td>
<td>9.5</td>
<td>88.5</td>
</tr>
<tr>
<td>7.4</td>
<td>98.1</td>
<td>8.0</td>
<td>97.0</td>
<td>8.9</td>
<td>110.5</td>
<td>9.8</td>
<td>99.3</td>
</tr>
<tr>
<td>7.4</td>
<td>99.2</td>
<td>8.0</td>
<td>103.3</td>
<td>8.9</td>
<td>103.2</td>
<td>9.8</td>
<td>80.9</td>
</tr>
<tr>
<td>7.4</td>
<td>103.1</td>
<td>8.0</td>
<td>112.5</td>
<td>8.9</td>
<td>95.4</td>
<td>9.8</td>
<td>79.3</td>
</tr>
<tr>
<td>7.0</td>
<td>107.2</td>
<td>8.1</td>
<td>108.6</td>
<td>8.9</td>
<td>92.8</td>
<td>9.8</td>
<td>79.0</td>
</tr>
<tr>
<td>7.7</td>
<td>104.3</td>
<td>8.3</td>
<td>105.1</td>
<td>8.9</td>
<td>91.9</td>
<td>10.1</td>
<td>72.8</td>
</tr>
<tr>
<td>7.7</td>
<td>100.6</td>
<td>8.3</td>
<td>89.9</td>
<td>8.9</td>
<td>85.6</td>
<td>10.1</td>
<td>87.0</td>
</tr>
<tr>
<td>7.7</td>
<td>96.8</td>
<td>8.6</td>
<td>71.6</td>
<td>9.2</td>
<td>86.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.7</td>
<td>96.4</td>
<td>8.6</td>
<td>93.0</td>
<td>9.2</td>
<td>89.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Write the null and alternative hypotheses for this study in words (use the term “association”).

In Section 10.2 we used the Corr/Regression applet to perform a simulation-based test for a correlation coefficient. Doing the simulation-based test for the slope of the regression line is extremely similar. The only difference is we use the slope as our statistic instead of the correlation coefficient. The applet has a radio button to select to switch from one statistic to the other.

2. Paste the data, handwidth, in the Corr/Regression applet and make a scatterplot. Describe the direction, form, and strength of the scatterplot.

3. Use the applet to determine the least squares regression line for predicting estimated weight based on hand width. What is the value of the slope of the regression line? What does this number imply in terms of hand width and estimated weight?

4. You should have found a negative association between hand width and estimated weight of the bean bag in the sample. The question, however, is if there were no association between hand width and the weight of an object in the population, how likely is it that we would get a slope as small as we did. Let’s apply the 3S Strategy.

   a. **Statistic:** What is the value of the slope in the sample?

   b. **Simulate:** To simulate you can use the same general approach we used for correlation in Section 10.2. Explain how you would conduct the approach by hand. Assume you have 46 slips of paper with the sample hand widths written on them, and 46 slips of paper with the perceived weight values written on them.

   c. **Strength of evidence:** Explain how you will calculate the strength of evidence in support of the conjecture that people with larger hands tend to perceive more weight in the bean bags.

5. a. Let’s shuffle the response variable (estimated weight) and see what we get for sample slope values when we break the association between the data pairs. Use the applet to do this five times and write down the five simulated slopes you get. Are any of them as small or smaller than the value of the actual sample slope?
b. Now do at least 1000 shuffles with the applet and find the p-value.

c. Click on Plot in the Shuffle Options section of the applet. This will now display each of the 1000 simulated regression lines (in blue). Describe the behavior of the 1000 regression lines across the different shuffles (see FAQ 10.4.1. for some discussion about this).

d. Explain what your p-value measures in context of the study.

e. Can we conclude that there is strong evidence of a genuine negative association between hand width and estimated weight of the object in the population?

6. As you did in Section 10.2, find the p-value corresponding to the correlation coefficient. How does this p-value compare with the p-value corresponding to the slope? (Note: You should have the p-value for the slope in #5d.)

**Key idea:** For a given dataset, the test for slope is equivalent to the test for correlation coefficient.

7. The sample used here was not a random sample. The article just said that the researchers obtained data on 46 participants.

   a. Describe the population to which we could legitimately make our inference. Explain your reasoning.

   b. Can we conclude that having smaller hands causes a person to estimate the weight of the bean bag to be smaller? Explain your answer.
FAQ 10.4.1 The point of averages

Q: I’ve got a question about the capitalized regression applet. When I shuffle to simulate the null hypothesis, all of the resulting ‘shuffled’ lines go through the same point. How come?

A: That point is sometimes called the “point of averages” because its coordinates are $\bar{x}, \bar{y}$. There’s more.

Q: Yes?

A: The point of averages is the center of gravity or balance point of the scatterplot. Imagine trying to find the point where the scatterplot balances. That point is always $\bar{x}, \bar{y}$.

Q: But that doesn’t answer my question. Why do the lines always go through that point?

A: When you shuffle, the $x$’s and $y$’s don’t change, so $\bar{x}$ and $\bar{y}$ don’t change.

Q: More please!

A: Here’s the key idea: Regardless of how the $x$’s and $y$’s are paired up, the least squares regression line will go through the point of balance $\bar{x}, \bar{y}$. Because shuffling doesn’t change the $x$ and the $y$ values the ‘shuffled’ lines all go through the same point.

Q: Why does the least-squares regression line go through there in the first place?

A: There’s some pretty complicated math in the full answer. The short answer is that the best way to minimize the squared residuals is to incorporate the balance point in the regression equation. Mathematically you can show that’s always the case.

Q: Does this have anything to do with using the mean as a measure of center?

A: Yes! The mean is the balance point when you have just one variable, and minimizes the sum of the squared deviations of the observations in your sample.
Section 10.4 Summary

The 3S Strategy can be used with the sample slope as the statistic of interest and applying the same simulation strategy we have used before.

- We can use the shuffling of the values of the response variable, to produce simulated values of the statistic that could have been if there was no genuine association between the two variables.
  - As always, the p-value is calculated as the proportion of repetitions in which the simulated value of the statistic is at least as extreme as the observed value of the statistic; and, a small p-value provides evidence against the null hypothesis.
- For a given dataset, the test of significance for the slope is equivalent to the test of significance for the correlation coefficient.

Section 10.5: Inference for the Regression Slope: Theory-Based Approach

Introduction

You may have noticed a very familiar looking pattern to the null distributions you generated of sample correlation coefficients in Section 10.2 and of sample slopes in Section 10.4. For example, two null distributions from Examples 10.2 and 10.4 are shown again in Figure 10.12.

Figure 10.12: Null distributions for sample correlation (left) and sample slope (right) from previous examples

Example 10.2: Exercise and Mood Intensity

Example 10.4: GPA and nonacademic hrs.
The null distribution on the left in Figure 10.11 is of sample correlations from Example 10.2 (change in mood and exercise intensity), and the null distribution on the right is of sample slopes from Example 10.4 (GPA and hours spent on nonacademic activities). Both are centered around zero and fairly bell-shaped.

Although we won’t precisely predict the correlation or slope null distributions we obtained through simulation, under certain conditions, the null distributions of a standardized version of these statistics are well modeled by $t$-distributions. Thus, theory-based inference can be conducted. This theory-based test has some additional differences from the simulation version which we will also explore.

**Example 10.5A: Predicting Heart Rate from Body Temperature**

We will use data on the heart rates and body temperature of healthy adults to see whether there is an association between body temperature and heart rate. The dataset consists of body temperatures and heart rates from 65 females and 65 males.

Let’s start by looking at a scatterplot (see Figure 10.13) of the heart rate (beats per minute) and body temperature (degrees Fahrenheit) for the 130 subjects.

**Think about it:** What does the scatterplot reveal about the relationship between heart rate and body temperature for this sample?

**Figure 10.13:** A scatterplot of the body temperature and heart rate of 130 individuals

<table>
<thead>
<tr>
<th>Sample data:</th>
<th>(explanatory, response)</th>
</tr>
</thead>
<tbody>
<tr>
<td>88 81</td>
<td>99.1 80</td>
</tr>
<tr>
<td>99.1 74</td>
<td>99.2 77</td>
</tr>
<tr>
<td>66 77</td>
<td>68 68</td>
</tr>
<tr>
<td>77 79</td>
<td>78 79</td>
</tr>
<tr>
<td>77 78</td>
<td>78 78</td>
</tr>
<tr>
<td>130</td>
<td></td>
</tr>
</tbody>
</table>

n = 130
From the scatterplot it appears there is a fairly weak positive linear association between these two variables. We also found the correlation coefficient and regression equation for predicting heart rate from body temperature by using the applet. The correlation coefficient is 0.257 and the regression equation is

\[ \text{Predicted Heart Rate} = -166.3 + 2.44(\text{Body Temp}). \]

**Think about it:** Interpret the slope value from this equation.

From the regression equation, we can see the slope is 2.44. This means that we predict heart rate to increase 2.44 beats per minute for every 1 degree rise in body temperature.

To evaluate whether this is convincing evidence that there is in fact a relationship between heart rate and body temperature in the population, we will first use a simulation-based approach. Our hypotheses are as follows:

- **Null Hypothesis:** There is no association between heart rate and body temperature in the population.

- **Alternative Hypothesis:** There is an association between heart rate and body temperature in the population.

The response variable (heart rate) in the dataset of 130 individuals was shuffled and redistributed to the body temperature values 10,000 times, and each time a simulated value for the slope of the regression line was computed. A graph of those simulated slopes is shown in Figure 10.14. We see that our observed sample slope of 2.44 is very unlikely in the null distribution. Only 36 out of 10,000 shuffles yielded a slope that was as extreme as 2.44 (either greater than or equal to 2.44 or less than or equal to -2.44). Thus our p-value of 0.0036 gives us strong evidence of a relationship between heart rate and body temperature in the population.
Figure 10.14: A null distribution for slope between heart rate and body temperature along with other output from the Corr/Regression applet

As you can see from Figure 10.14, the null distribution has the nice, familiar bell shape. However, as has been the case before, we need a standardized version of the test statistic in order to obtain an accurate theory-based prediction of the null distribution.

Formulas

The theory-based approach computes a standardized statistic ($t$-statistic) using either of the following equations:

$$ t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} $$

More details are provided in the Calculation details appendix, but notice that the $t$-statistic can be computed based on either the correlation coefficient or the slope, and it yields the same standardized value—further underscoring that tests for the correlation and for the slope are essentially identical.
Figure 10.15 shows the results of simulating when using the \( t \)-statistic.

**Figure 10.15:** Simulated standardized statistics with \( t \)-distribution overlaid

Validity Conditions

As has been the case before, theory-based tests have validity conditions which are needed in order to obtain an accurate prediction of the simulated distribution.

**Validity conditions for a theory-based test for regression**

1. The general pattern of the points in the scatterplot should follow a linear trend; the pattern should not show curved or other non-linear patterns
2. There should be approximately the same distribution of points above the regression line as below the regression line (symmetry about the regression line) and,
3. The variability of the points around the regression line should be similar regardless of the value of the explanatory variable; the variability (spread) of the points around the regression line should not differ as you slide along the \( x \)-axis (equal variance/standard deviation).
To check the validity conditions for using the theory-based approach, we examine the scatterplot in more detail. First, the pattern of the points in the scatterplot doesn’t looked curved; a linear trend seems reasonable. Second, we note that the points above the regression line appear similar in spread and shape as the points below the regression line (like a mirror image to the pattern). Finally, the variability in heart rates is approximately the same for different body temperatures (e.g., we don’t see that low body temperatures have low variability in heart rates, and high body have high variability in heart rates, or vice versa). When checking validity conditions, we use logic much like we do when testing hypotheses—we will assume the condition is true unless we see strong evidence that it is not.

You know by now that theory-based tests, when valid, predict the behavior of the null distribution (of the standardized statistic) you would have gotten from simulation, if you had simulated. In this case, we know that the theory-based approach will accurately predict the shape of the null distribution because the validity conditions are met. This is confirmed by overlaying the theoretical $t$-distribution on our simulated slopes (Figure 10.15). But, it’s actually a bit more complicated than that as FAQ 10.5.1 and the discussion in the next sections point out.

Figure 10.16 shows a more formal presentation of the output for a regression analysis illustrating some of the calculation details.

**Figure 10.16: Formal Regression Table output for theory-based approach**

<table>
<thead>
<tr>
<th>Term</th>
<th>Coeff</th>
<th>SE</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-166.28</td>
<td>80.91</td>
<td>-2.06</td>
<td>0.0419</td>
</tr>
<tr>
<td>BodyTemp</td>
<td>2.44</td>
<td>0.82</td>
<td>2.97</td>
<td>0.0036</td>
</tr>
</tbody>
</table>

The last line of the regression output shows the theory-based p-value for our test. This p-value of 0.0036 is the same as our simulated p-value of 0.0036. The theory-based p-value will always be given as a two-sided p-value in the Regression Table (see Figure 10.16). In this case, we are conducting a two-sided test. If we had suspected beforehand that heart rate and body temperatures have a positive association, then our p-value would simply be half of the theory-based p-value of 0.0036, namely 0.0018.

Also note that the output reports the standardized statistic of $t = 2.97$ for the slope. This tells us that the observed sample slope (2.44) is almost 3 standard errors (0.82) above the hypothesized slope of zero. Note that the standard error here is telling us about the shuffle-to-shuffle variation in the sample slopes. This value is smaller than what we saw in the simulation for the standard deviation of the simulated slopes (there SD of simulated statistics was 0.842 (see Figure 10.14)) because (as discussed in FAQ 10.5.1) the simulation is for a more general ‘no association’ hypothesis, whereas the theory-based p-value
is for a more specific set of hypotheses about linearity. In particular we can now specify that we are testing for linearity,

Null Hypothesis: There is no *linear* association between heart rate and body temperature in the population.

Alternative Hypothesis: There is a *linear* association between heart rate and body temperature in the population.

This gives us the ability to state hypotheses using population parameters. We use the symbol $\beta$ (beta) to indicate the population slope and $\rho$ (rho) to indicate the population correlation. We can write the hypotheses as shown here:

Null Hypothesis: $\beta = 0$ (the population slope = 0)
Alternative Hypothesis: $\beta \neq 0$ (the population slope is not 0)

Or, equivalently,

Null Hypothesis: $\rho = 0$ (the population correlation = 0)
Alternative Hypothesis: $\rho \neq 0$ (the population correlation is not 0)

**Confidence interval**

The theory-based confidence interval (Figure 10.17) for the slope tells us that we are 95% confident that a one degree increase in body temperature is associated with an average increase of 0.8137 to 4.0728 heart beats per minute in the population represented by this sample. Note that this can be approximated using the 2SD method using the more precise estimate of the standard deviation from the regression table (0.82), yielding $2.44 \pm 1.64 = (0.80, 4.08)$. See FAQ 10.5.1 for more details.
FAQ 10.5.1 Simulation, parameters, and intervals

Q: I haven’t forgotten about the draft lottery. You still owe me more about parameters and intervals.
A: We’re now in a better position to take a closer look. How would you ask your question based on what you’ve seen since Section 10.2?
Q: Back then we did a simulation of the null hypothesis of no association by shuffling the response variable values. We had no parameter, and without a parameter, there was no way to get a confidence interval.
A: Right.
Q: Then in the very next section parameters popped up like mushrooms after a rain. What changed?
A: That’s definitely a thoughtful question. You’ve put your finger on a very important idea. Remember that with the lottery you said you couldn’t tell what the pattern of possible association might look like, and I agreed.
Q: OK.
A: Now look back and compare the scatterplot for the lottery (Figure 10.6) and the scatterplot for the dinner plates (Figure 10.7).
Q: What’s your point?
A: For the dinner plates, the form of the plot is roughly linear, and we are most interested in the underlying linear relationship.
Q: OK again.
A: We can make that linearity part of our framework for inference.
Q: Framework?
A: We make linearity one of our validity conditions.
Q: Why would I want to do that?
A: If the relationship is linear, we get parameters: the slope of the line, the intercept, and the correlation coefficient. These parameters don’t make sense unless the relationship is linear.
Q: If I understand you, we have two kinds of situations. For the first kind, like the draft lottery, we know the null hypothesis – no association – but we don’t specify anything more specific than “is an association” for the alternative hypothesis. We can test the statistical significance of the observed association by shuffling, but we don’t have a parameter, so we don’t have intervals.
A: Excellent! And the other kind?
Q: We assume that the underlying relationship is linear. We have parameters, and we can find theory-
based tests and intervals.
A: Well said.
Q: But I’ve still got a question.
A: Sure.
Q: Why can’t we do a simulation-based significance test for the dinner plates? Doesn’t it work when the alternative is a linear relationship?
A: You could do it, and it does work, but the theory-based test is more powerful.
Q: How come?
A: If you assume linearity, you automatically rule out any and all non-linear forms of association. You narrow your question to whether the population slope is zero. Nothing else matters.
Q: That makes sense.
A: One way you can see that it’s more powerful is by looking at the variability in the null distribution. The standard deviation of the null distribution of slopes when we simulate using shuffling is larger than the standard deviation of the sample slopes using the theory-based test.
Q: …because of the extra assumption of being linear. That makes sense.
A: Here’s another way to look at it. You go to the doctor. The null hypothesis is that don’t have a disease. The alternative hypothesis is that do have some disease. What’s the parameter?
Q: I get it. No way to define a parameter without additional assumptions.
A: Now suppose the question is much more narrow: whether you have diabetes. The doctor knows to do a test based on the level of your fasting blood sugar.
Q: The parameter.
A: Exactly.
Q: One last question. Are you telling me that we can’t do simulation-based confidence intervals for the slope?
A: I’m glad you asked that, because you actually can get simulation-based intervals if you have a linear relationship. It’s a little more complicated, so we’ve saved it for your next course in statistics.

---

**Example 10.5B: Smoking and Drinking**

The scatterplot in Figure 10.18 shows the relationship between number of alcoholic drinks consumed per week (x-axis) and number of cigarettes smoked per week (y-axis) for a random sample of students at a small college.

Note: The scatterplot will only show one “dot” for people who have the same values of the explanatory/response variable. So, although we only see one “dot” on the point (0, 0) there are actually 524 students at that point!
Figure 10.18: A scatterplot of alcoholic drinks and cigarettes smoked per week for a sample of college students

Think about it: Does the data in the scatterplot appear to meet the validity conditions to conduct a theory-based test?

Examining the scatterplot of the sample data should help reveal any strong violations of the validity conditions described in Example 10.5A. Both conditions appear violated. In particular, the scatterplot in Figure 10.18 reveals that most dots lying below the regression line (along the x-axis) are much closer to the line on average than the values lying above the regression line. For example, at any value of number of drinks per week, there are very many people who smoked zero cigarettes, a few with a small number of cigarettes smoked, and then even fewer with very few large values for number of cigarettes smoked. In other words, the distribution of cigarettes smoked per week for each different value of alcoholic drinks per week is highly right skewed. The other condition is that there is similar variability of the dots around the regression line for different values of the explanatory variable (alcoholic drinks per week). The scatterplot suggests that there is evidence that this condition is also violated because the distributions of the number of cigarettes smoked seem to be more variable for larger values of alcoholic drinks per week than for smaller values.

Because these two conditions are violated, we should not apply theory-based inference procedures to these data. In Figure 10.19 we see both the theory-based prediction for the null distribution of the standardized correlation coefficient and simulation results for 1000 shuffles. Notice that the theory-based
prediction of the null distribution is mound-shaped (a $t$ distribution) with a mean of zero, but the simulated null distribution for the standardized correlation coefficient is highly right-skewed. In this case, the theory-based prediction for the null distribution is not a good match to the actual null distribution, so p-values and confidence intervals from the theory-based approach could be quite misleading.

**Figure 10.19:** Theory-based and simulated null distribution for $t$-statistics by standardizing shuffled slopes of alcohol drinks and cigarettes smoked per week

When the validity conditions are not met for theory-based inference for a population correlation coefficient or slope, you can rely on the simulated-based approach. Another common strategy is to re-express ("transform") the data on a different scale, such as a logarithmic scale, in such a way that the conditions would be met. See Exercise 10.5.20 for an example.

**Caution: Outliers and Influential Observations**

Reconsidering Figure 10.18, we also note that there may be some influential points in this sample. Specifically, the person who says they smoke 130 cigarettes a week is also one of largest alcohol consumers (50 drinks per week). If we remove this person from the dataset, the correlation coefficient drops from 0.37 to 0.30, a noticeable decrease. This is the best way to check whether extreme observations unduly affect the analysis: Remove them from the dataset and see how the relationship changes. It is a judgment call about how big of a change is large, but any point that substantially changes the correlation coefficient or regression slope is considered influential. Depending on the context, influential points should be considered for removal from the analysis in order to get a more robust estimate of the correlation coefficient and regression line, but should only be done so with proper justification.
**Key idea:** Remove possible influential points and outliers and recompute the correlation coefficient and/or regression slope. If the statistic has changed a lot after removing the point, the point is considered influential and you may want to keep that point out of the analysis (or report your conclusions both with and without that observation).

---

**Exploration 10.5: Predicting Brain Density from Number of Facebook Friends**

How often did you check your Facebook account today? How many Facebook friends do you have? Does everyone have the same number? Might the number of facebook friends that a person has be associated in some way with the person’s brain structure? Researchers used MRI’s to examine areas of the brain that are involved with social interaction, memory, and emotional responses (Kanai, Bahrami, Roylance, & Rees, 2011). Their subjects were 40 university students in London. They examined five areas of the brain that have been previously linked to social perception and associative memory, and they compared each to the number of Facebook friends the subject reported. The results from each brain area were quite similar. We will look at one set of measurements focusing on the left MTG (middle temporal gyrus which has been linked in other studies to face recognition). These results are shown in Table 10.5. Note that number of friends is given in units of 100 friends, so that 0.30=30 friends, 1.09=109 friends, and so on.

**Table 10.5:** The number of Facebook friends (in 100s of friends) and the brain density (in arbitrary units) for 40 university students

<table>
<thead>
<tr>
<th>friends</th>
<th>density</th>
<th>friends</th>
<th>density</th>
<th>friends</th>
<th>Density</th>
<th>friends</th>
<th>density</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.30</td>
<td>-2.14</td>
<td>3.41</td>
<td>-0.72</td>
<td>5.08</td>
<td>-0.17</td>
<td>2.93</td>
<td>1.24</td>
</tr>
<tr>
<td>1.09</td>
<td>-1.09</td>
<td>3.97</td>
<td>-0.74</td>
<td>5.15</td>
<td>-0.09</td>
<td>3.09</td>
<td>0.93</td>
</tr>
<tr>
<td>0.39</td>
<td>-0.72</td>
<td>4.10</td>
<td>-0.85</td>
<td>5.01</td>
<td>-0.03</td>
<td>3.65</td>
<td>1.02</td>
</tr>
<tr>
<td>1.20</td>
<td>-0.63</td>
<td>4.39</td>
<td>-0.69</td>
<td>3.92</td>
<td>0.26</td>
<td>4.73</td>
<td>0.75</td>
</tr>
<tr>
<td>0.80</td>
<td>-0.43</td>
<td>4.60</td>
<td>-0.71</td>
<td>3.13</td>
<td>0.42</td>
<td>3.95</td>
<td>1.34</td>
</tr>
<tr>
<td>1.21</td>
<td>-0.26</td>
<td>5.07</td>
<td>-1.07</td>
<td>2.53</td>
<td>0.46</td>
<td>6.14</td>
<td>0.96</td>
</tr>
<tr>
<td>2.24</td>
<td>-0.34</td>
<td>5.08</td>
<td>-0.49</td>
<td>1.51</td>
<td>0.23</td>
<td>7.32</td>
<td>1.05</td>
</tr>
<tr>
<td>2.40</td>
<td>-1.51</td>
<td>5.83</td>
<td>-0.59</td>
<td>1.73</td>
<td>0.27</td>
<td>6.17</td>
<td>1.6</td>
</tr>
<tr>
<td>3.23</td>
<td>-1.45</td>
<td>6.34</td>
<td>-0.41</td>
<td>1.52</td>
<td>0.55</td>
<td>5.47</td>
<td>1.97</td>
</tr>
<tr>
<td>3.81</td>
<td>-1.48</td>
<td>5.63</td>
<td>0.21</td>
<td>2.71</td>
<td>1.09</td>
<td>5.70</td>
<td>2.08</td>
</tr>
</tbody>
</table>

The researchers wanted to explore whether there is a positive relationship between these two variables, and we will do the same. In particular, the researchers tried to predict brain density based on number of facebook friends, asking: Does brain density tend to increase as a person’s number of facebook friends increases?
1. What are the null and alternative hypotheses for this study?

2. Paste the data (Facebook) in the Corr/Regression applet and make a scatterplot where brain density is the response variable and number of Facebook friends is the explanatory variable. (Note that the number of facebook friends is reported in terms of hundreds of friends.)

Describe the direction, form, and strength of association between the variables as revealed in the scatterplot. Are there any unusual observations?

3. What is the slope of the regression equation for predicting brain density based on number of Facebook friends? How does this value support your answer to the previous question?

4. You should have found a positive slope between brain density and number of Facebook friends. The question, however, is if there were no association between brain density and number of Facebook friends, how likely is it that we would get a slope as high as we did.

   a. In your own words, explain how you could use slips of paper to model the null hypothesis of no association between brain density and number of Facebook friends.

   b. Let’s shuffle the response variable (brain density) to see what we get for the slope if we break the association between the data pairs. Use the applet to do this five times and write down the
five simulated slopes you get. Are any of them as large as or larger than the actual value of the sample slope?

c. Now do at least 1000 shuffles with the applet and find your p-value.

d. Interpret what your p-value measures in context of the study.

e. Can we conclude that there is a positive association between the density of the left MTG and the number of Facebook friends a person has in the population? Explain your reasoning.

5. What is the shape of the null distribution of shuffled slopes?

By now, you’ve seen many times that when the null distribution of statistics takes a familiar, mound-shaped curve, we can often use theory-based methods to predict the null distribution of related standardized statistics---as long as certain validity conditions are true. This is no different for regression (and correlation!).

The theory-based approach computes a standardized statistic ($t$-statistic) using one of the following equations:

$$ t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} $$

More details are provided in the Calculation details appendix, but notice that the $t$-statistic can be computed based on either the correlation coefficient or the slope, and it yields the same value-further underscoring that tests for the correlation coefficient and for the slope are essentially identical.

6. Use the applet to find the correlation coefficient.
7. Use the correlation coefficient to find the t-statistic using the equation shown above.

There are three validity conditions for regression which are needed in order to use the theory-based approach to yield a p-value

<table>
<thead>
<tr>
<th>Validity conditions for a theory-based test for regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The general pattern of the points in the scatterplot should follow a linear trend; the pattern should not show curved or other non-linear patterns,</td>
</tr>
<tr>
<td>2. There should be approximately the same distribution of points above the regression line as below the regression line (symmetry about the regression line) and,</td>
</tr>
<tr>
<td>3. The variability of the points around the regression line should be similar regardless of the value of the explanatory variable; the variability (spread) of the points around the regression line should not differ as you slide along the x-axis (equal variance/standard deviation).</td>
</tr>
</tbody>
</table>

8. Based on the scatterplot you used in #2, is Validity Condition 1 met? Namely, is the general pattern of the scatter plot linear?

9. Based on the scatterplot you used in #2, is Validity Condition 2 met? Namely, is the distribution of points above the regression line the same as below the line?

10. Based on the scatterplot you used in #2, is Validity Condition 3 met? Namely, is the variability of the points around the line similar regardless of the value of the explanatory variable?

11. Select the t-statistic radio button, then check the Overlay t distribution button. Does the t-distribution appear to do a good job of predicting the distribution of the simulated t-statistics? Why?
12. In the applet, click the box for **Regression Table**. This table provides the observed $t$-statistic for the data, as well as the two-sided, theory-based test $p$-value. Use the theory-based approach to test whether there is a significant association between the variables in the population. Report the standardized statistic and $p$-value. Summarize your conclusion.

When using the theory-based approach for regression you can write your hypotheses in terms of population parameters. The relevant population parameters are the population slope (indicated by the Greek letter $\beta$) and the population correlation (indicated by the Greek letter $\rho$). See FAQ 10.5.1 for more discussion about parameters.

13. Write the null and alternative hypotheses in terms of the population slope.

14. Write the null and alternative hypotheses in terms of the population correlation.

15. Produce and interpret a 95% confidence interval for the population slope coefficient. *(Hint: Make sure that your interpretation refers to how to interpret the slope coefficient in the population and remember that the units of the explanatory variable is ‘100s of facebook friends’.) To get the confidence interval, check the box below the regression table for 95% Confidence interval for Slope.

16. Imagine if we added someone to this dataset whose brain density was -10 and had 5000 Facebook friends. Uncheck the Show Shuffle Options and enter this observation into the dataset using **Add/remove observations** or typing this new row into the Sample data window. Conduct a two-sided test. Caution: Make sure you enter 5000 facebook friends in terms of the units used in this study. What would 5000 facebook friends be in terms of ‘100s of facebook friends’?

a. Find the $p$-value using a simulation-based approach with shuffled $t$-statistics.

b. Find the $p$-value using the theory-based approach.

c. Why are the $p$-values from the two approaches not very similar?
17. Remember that the sample used here was not randomly selected, but rather a group of 40 volunteer university students.

   a. Describe a population in which you would be comfortable drawing inferences. Explain your reasoning.

   b. Can we conclude that the acquiring more Facebook friends would lead to an increase in a person’s brain density? Explain your answer.

   c. What about the opposite direction: Can we conclude that having a larger brain density causes a person to acquire more Facebook friends? Explain.

18. The researchers originally searched for positive associations between many areas of the brain and the number of Facebook friends for 125 subjects. They found significantly positive associations for a number of brain areas. The second part of the study is with the 40 subjects that we looked at in this exploration.

   Why do you think this method of ”look at a bunch of regions to see which are significant” is often criticized? (Hint: If you run lots of tests of significance using a 5% level of significance, how often will you reject the null hypothesis when you really shouldn't have?)

**Section 10.5 Summary**

A theory-based method can be used to conduct inference for the population slope coefficient or population correlation coefficient. We consider the two methods identical.

- The test of significance and the confidence interval are based on the \( t \)-distribution.
- The validity conditions for the slope coefficient are that the response variable has approximately the same distribution of points above the regression line as below the regression line (symmetry) and, and the variability of the points around the regression line should be similar regardless of the value of the explanatory variable (equal variance).
Chapter 10 Summary

In this chapter we looked at relationships between two quantitative variables. We saw that scatterplots are used to explore the relationship between two quantitative variables. In creating a scatterplot, we put the explanatory variable on the horizontal axis and the response variable on the vertical axis. Which variable is which is sometimes straightforward (especially if you are doing an experiment) and sometimes not so straightforward.

We saw that the correlation coefficient is a measure of the strength and direction of a linear relationship between two quantitative variables.

A regression line has the form \( \hat{y} = a + bx \) where \( a \) is the \( y \)-intercept, \( b \) is the slope, \( x \) is a value of the explanatory variable, and \( \hat{y} \) is the predicted value for the response variable. You should always put the variables in context when writing out the equation and be able to give contextual interpretations of the slope and intercept (but watch for extrapolation). For a specific value of \( x \), the corresponding \( y - \hat{y} \) is a residual. The least squares regression line is found by minimizing the sum of the squared residuals.

The simulation-based technique that we learned in the previous chapters was adapted to work with looking for an association in quantitative data. We again used the 3S Strategy:

1. **Statistic:** In this chapter our statistic was either the correlation coefficient or the regression slope.

2. **Simulate:** Shuffle the response variable, recompute the statistic, and repeat this process many times.

3. **Strength of evidence:** We find evidence against the null hypothesis if the original sample statistic is in the tail of the distribution used to model the null hypothesis.

Similar to previous chapters, we saw that the p-value is the probability of getting a value of the (standardized) statistic equal to or more extreme than the one from our original sample data when the null hypothesis is true. We also used the standard deviation from the null distribution and the two standard deviation rule to approximate a 95% confidence interval for the parameter.

The theory-based test for a regression slope was discussed and compared to the simulation-based test. We saw how the predicted null distribution was similar to all the others we have seen so far: symmetric, bell-shaped, and centered at zero. This allowed us (the computer) to formally standardized the statistic and use the \( t \)-distribution (with \( n - 2 \) degrees of freedom) to approximate the p-value and estimate the confidence interval.
Chapter 10 Glossary

correlation coefficient
Statistic that measures direction and strength of a linear relationship between two quantitative variables
extrapolation
Predicting values for the response variable for given values of the explanatory variable that are outside of the range of the original data
influential observations
An observational is considered influential if removing it from the data set dramatically changes the correlation coefficient or regression line. Often have extreme x values
predictor
Another word for explanatory variable, often used in correlation/regression settings
r
Symbol for correlation coefficient, values range from −1 to 1 and are unit-less. Values close to −1 and 1 denote a strong linear relationship whereas values close to 0 denote a weak or no linear relationship
residuals
The vertical distances between a point and the least squares regression line
slope
Change in predicted response variable divided by change in explanatory variable
SSE
Sum squared error. It is the sum of all the squared residuals
transform
Express data on a different scale, such as logarithmic, often used to meet validity conditions
unusual observations
Any observations that do not fit in with the pattern of the rest of the observations
validity conditions for regression theory-based test
Symmetry of data about the regression line and equal variability in response variable across the x-values
**y-intercept:** ................................................................. 10-38

β
Symbol for the population slope .................................................. 10-58

ρ
Symbol for the population correlation ......................................... 10-58
Exercises

Section 10.1

10.1.1* Suppose that a study measures how far each student sits from the front of the classroom and also records the student’s final exam score. If better students tend to sit closer to the front, would the correlation between distance and exam score be positive, negative, or close to zero? Explain briefly.

10.1.2 Suppose that you record the daily high temperature and the daily amount of ice cream sold by an ice cream vendor at your favorite beach this summer, starting on Memorial Day weekend and ending on Labor Day weekend. Would you expect to find a positive or negative correlation between these variables? Explain briefly.

10.1.3* Suppose that every student in this class scored 5 points lower on the second exam than on the first exam. Consider the correlation coefficient between first exam score and second exam score. What would the value of this correlation coefficient be? Explain briefly. (Hint: You might draw a scatterplot of hypothetical data that fit the description.)

10.1.4 Six statements about what you can learn from the correlation coefficient are listed below. Label each statement, choosing from A (can always tell), N (can never tell), or S (can sometimes tell. For each statement you label S for sometimes, explain briefly when you can tell and when you cannot.

   a) Whether there is a systematic relationship between \( x \) and \( y \).
   b) Whether the relationship between \( x \) and \( y \) is linear or curved.
   c) Whether the least squares line has a positive or negative slope.
   d) How closely the points fall near a line.
   e) Whether there are outliers or influential points.
   f) How steep the least squares line is.

10.1.5* Fill in the blanks, choosing from: values of the response, values of the explanatory variable, cases:
In a scatterplot, the points are ________________; the \( x \)-axis represents ________________; and the \( y \)-axis represents ________________.

10.1.6 The data shown in the following scatterplot shows a very nice relationship between the two variables. However, the correlation here is 0.03, very close to zero. Explain why we can have a nice relationship between two quantitative variables and yet have a correlation of 0.

10.1.7* Which of the following is not a property of correlation, \( r \)?

A. \(-1 \leq r \leq 1\)
B. Correlation measures the strength of a linear relationship between two quantitative variables.
C. The sign on \( r \) tells the direction of the linear relationship between two quantitative variables.
D. If the correlation between two quantitative variables is zero, then there is no relationship between these two variables.

10.1.8 Which of the following statements is correct?

A. Changing the units of measurements of the explanatory or response variable does not change the value of the correlation.
B. A negative value for the correlation indicates that there is no relationship between the two variables.
C. The correlation has the same units (e.g., feet or minutes) as the explanatory variable.
D. Correlation between \( y \) and \( x \) has the same number but opposite sign as the correlation between \( x \) and \( y \).
10.1.9* If two variables are negatively associated, then we know that
A. Above average values in one variable correspond to below average values in the other variable.
B. Above average values in one variable correspond to above average values in the other variable.
C. Below average values in one variable correspond to below average values in the other variable.
D. Below average values in one variable correspond either above average or below average values in the other variable.

10.1.10 For each of the following statements, say what, if anything is wrong.

a) Because the correlation coefficient between test time and test score is -0.56, the correlation coefficient between test score and test time must be 0.56 (the positive value).
b) There is a strong positive correlation between a person’s yard size and whether or not they have a dog.
c) For a sample of 50 students, the correlation coefficient between weight (kg) and height (inches) was found to be 0.78kg/inches.
d) A correlation coefficient of $r = 0.84$ denotes just as strong a relationship as a correlation coefficient of $r = -0.84$.

House Prices*

10.1.11 The data file HousePrices contains data on prices ($) and sizes (in square feet) for a random sample of houses that sold in the year 2006 in Arroyo Grande, California

a) Identify the type of study randomized experiment or observational study.
b) Identify the experimental/observational units in this study.
c) Identify the two variables of interest, and whether each is categorical or quantitative.

Which variable do you think makes more sense to use as the explanatory variable, and which as the response variable?
d) Enter the data into the Corr/Regression applet. Produce a scatterplot, and comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)
e) Report the value of the correlation coefficient. Does the value of the correlation coefficient support your answer in part d)? Explain how you are deciding.

Roller coasters

10.1.12 The data file RollerCoasters contains data (collected from www.rcdb.com) on the maximum speed (miles per hour) and the maximum height (feet) for each of a sample of roller coasters in the U.S.

a) Identify the type of study: randomized experiment or observational study.
b) Identify the experimental/observational units in this study.
c) Identify the two variables of interest, and whether each is categorical or quantitative. Which variable do you think makes more sense to use as the explanatory variable, and which as the response variable?
d) Enter the data into the Corr/Regression applet. Produce a scatterplot, and comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)
e) Report the value of the correlation coefficient. Does the value of the correlation coefficient support your answer in part d)? Explain how you are deciding.

Height and finger length*

10.1.13 An article titled “Giving the Finger to Dating Services” appeared in Chance magazine (2008) that investigated the association between length of index finger and height. This research was spurred by a report that online dating sites were asking members for the length of their index fingers. Perhaps this information could be used to verify whether the member was reporting his/her correct height on the profile? The data file HeightAndFingerLength contains data on the height (inches) and the length of index finger of dominant hand (cm) for students in a statistics class.

a) Identify the type of study: randomized experiment or observational study.
b) Identify the experimental/observational units in this study.
c) Identify the two variables of interest, and whether each is categorical or quantitative. Which variable do you think makes more sense to use as the explanatory variable, and which as the response variable?
d) Enter the data into the Corr/Regression applet. Produce a scatterplot, and comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)
e) Report the value of the correlation coefficient. Does the value of the correlation coefficient support your answer in part d)? Explain how you are deciding.

10.1.14 Refer to the previous exercise that investigated the association between length of index finger and height. The data file HeightAndFingerLength contains data on the height (inches) and the length of index finger of dominant hand (cm) for students in a statistics class.

a) Using appropriate technology (applet or statistical software) create a scatterplot of the data. On the scatterplot, locate the individual who reported a finger length of 5 cm and height of 70 inches. Does this individual appear to be an unusual observation? Explain you reasoning.
b) Next, report the correlation coefficient for these data.
c) Now, remove this individual (finger length of 5 cm and height of 70 inches). Recalculate the correlation coefficient for the data without this individual.
d) Did the correlation coefficient increase (move farther from zero) or decrease (move closer to zero) on the removal of the individual who reported a finger length of 5 cm and height of 70 inches? Explain why that makes sense.

10.1.15: Refer to the previous exercises that investigated the association between length of index finger and height. The data file HeightAndFingerLength contains data on the height (inches) and the length of index finger of dominant hand (cm) for students in a statistics class.

a) Once again, using appropriate technology (applet or statistical software) create a scatterplot of the data. On the scatterplot, locate the individual who reported a finger length of 10.16 cm and height of 78 inches. Does this individual appear to be an unusual observation? Explain you reasoning.
b) Next, report the correlation coefficient for these data.
c) Now, remove this individual (finger length of 10.16 cm and height of 78 inches). Recalculate the correlation coefficient for the data without this individual.
d) Did the correlation coefficient increase (move farther from zero) or decrease (move closer to zero) on the removal of the individual who reported a finger length of 10.16 cm and height of 78 inches? Explain why that makes sense.

**Relationship between parents’ and their children’s heights**

**10.1.16** In 1885, Sir Francis Galton first used regression to explore the association between children’s heights and (biological) parents’ heights. The data file **MomandChildHeights** contains data for students in a statistics class on the following two variables: student’s height (inches) and biological mother’s height (inches).

a) Identify the type of study: randomized experiment or observational study.
b) Identify the experimental/observational units in this study.
c) Identify the two variables of interest, and whether each is categorical or quantitative. Which variable do you think makes more sense to use as the explanatory variable, and which as the response variable?
d) Enter the data into the **Corr/Regression** applet. Produce a scatterplot, and comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)
e) Report the value of the correlation coefficient. Does the value of the correlation coefficient support your answer in part d)? Explain how you are deciding.

**10.1.17** Refer back to the previous exercise that explores the association between children’s heights and (biological) mothers’ heights. The data file **DadandChildHeights** contains data for students in a statistics class on the following two variables: student’s height (inches) and biological father’s height (inches).

a) Identify the type of study: randomized experiment or observational study.
b) Identify the experimental/observational units in this study.
c) Identify the two variables of interest, and whether each is categorical or quantitative. Which variable do you think makes more sense to use as the explanatory variable, and which as the response variable?
d) Enter the data into the *Corr/Regression* applet. Produce a scatterplot, and comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)

e) Report the value of the correlation coefficient. Does the value of the correlation coefficient support your answer in part d)? Explain how you are deciding.

10.1.18 Refer back to the previous two exercises that explore the association between children’s heights and (biological) parents’ heights. Recall that the data file *DadandChildHeights* contains data for students in a statistics class on the following two variables: student’s height (inches) and biological father’s height (inches), whereas the data file *MomandChildHeights* contains data for students in a statistics class on the following two variables: student’s height (inches) and biological mother’s height (inches).

a) Report the value of the correlation coefficient for children and fathers’ heights.
b) Report the value of the correlation coefficient for children and mothers’ heights.
c) In which case does there appear to be a stronger association between child and parent’s heights: with father’s height or with mother’s height? Explain your reasoning.

Scrabble names*

10.1.19 Have you ever played the game *Scrabble*? It is a word building game where certain letters earn you more points than others when you use them in your word. Here is the table of letters and the corresponding Scrabble points:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

In a statistics class, students were asked to calculate the number of Scrabble points their names would earn. For example, the Scrabble score for the name Tom Sawyer would be $1 + 1 + 3 + 1 + 1 + 4 + 4 + 1 + 1 = 17$ points. The data file *ScrabbleNames* contains data on the following two variables: number of letters in a student’s name and the corresponding Scrabble score.

a) Identify the type of study: randomized experiment or observational study.
b) Identify the experimental/observational units in this study.
c) Identify the two variables of interest, and whether each is categorical or quantitative. Which variable do you think makes more sense to use as the explanatory variable, and which as the response variable?

d) Enter the data into the Corr/Regression applet. Produce a scatterplot, and comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)

e) Report the value of the correlation coefficient. Does the value of the correlation coefficient support your answer in part d)? Explain how you are deciding.

10.1.20 Refer back to the previous exercise about exploring the relationship between the number of letters in a name and the corresponding Scrabble score. Recall that students were asked to calculate the number of Scrabble points their names would earn. Students were also asked to calculate the points per letter, that is, \( \text{ratio} = \frac{\text{Scrabble score}}{\text{number of letters in name}} \). The data file ScrabbleRatio contains Scrabble points for a student’s name and the corresponding ratio.

a) Enter the data into the Corr/Regression applet. Produce a scatterplot, and comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)

b) Report the value of the correlation coefficient. Does the value of the correlation coefficient support your answer in part a)? Explain how you are deciding.

10.1.21 Refer back to the previous two exercises that explore the association between number of letters in a name and the corresponding Scrabble score. Recall that the data file ScrabbleNames contains data on number of letters in students’ names and the corresponding Scrabble points, whereas the data file ScrabbleRatio contains Scrabble points for a student’s name and the corresponding ratio of Scrabble points to number of letters.

a) In which case is the linear association stronger: between the number of letters in a student’s name and the corresponding Scrabble points, or between Scrabble points for a student’s name and the corresponding ratio? How are you deciding?

b) With regard to your finding in part a), explain why that makes sense.
TV and life expectancy

10.1.22 The data file TVLife provides information on life expectancy and number of televisions per thousand people in a sample of 22 countries, as reported by the 2006 World Almanac and Book of Facts.

a) Identify the type of study: randomized experiment or observational study.

b) Identify the experimental/observational units in this study. (Hint: “People” is not the answer.)

c) Enter the data into the Corr/Regression applet. Produce a scatterplot using life expectancy as the response variable and number of TVs per 1000 as the explanatory variable. Comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)

d) Report the value of the correlation coefficient. Does the value of the correlation coefficient support your answer in part c)? Explain how you are deciding.

e) As you (should) have discovered in parts c) and d), the association between life expectancy and number of televisions per thousand people appears pretty strong. Based on this finding, is it okay to conclude that simply sending televisions to the countries with lower life expectancies would cause their inhabitants to live longer? Explain why or why not.

Section 10.2

10.2.1* An instructor wanted to investigate whether there was an association between height (inches) and hand span (cm). She collected data from 10 students, and after analyzing the data found the p-value for to be 0.022. For each of the following statements, indicate whether or not the statement is VALID or INVALID.

a) The p-value says that there is a 2.2% probability that there is no association between height and hand span.

b) The p-value says that there is a 2.2% probability that there is an association between height and hand span.
c) If there were no association between height and hand span, the probability of observing the association observed in the sample data of 10 students is 0.022.

d) If there were no association between height and hand span, the probability of observing the association observed in the sample data or an even stronger association in a sample of 10 students is 0.022.

e) If there were an association between height and hand span, the probability of observing the association observed in the sample data or an even stronger association in a sample of 10 students is 0.022.

10.2.2 A researcher wants to investigate whether there is a relationship between annual company profit ($) and median annual salary paid by the company ($). The researcher collects data on a random sample of companies, and after analyzing the data finds the p-value to be 0.56. Which of the following is an appropriate conclusion based on this p-value?

1. There is no relationship between annual company profit ($) and median annual salary paid by the company ($).

2. There is a relationship between annual company profit ($) and median annual salary paid by the company ($), and corresponding value of correlation coefficient is $r = 0.56$.

3. There is no evidence of a relationship between annual company profit ($) and median annual salary paid by the company ($).

4. There is a 56% chance that there is no relationship between annual company profit ($) and median annual salary paid by the company ($).

10.2.3* Two researchers want to investigate whether there is a relationship between annual company profit ($) and median annual salary paid by the company ($). Researcher Bart collects data on a random sample of 40 companies, and researcher Lisa collects data on a random sample of 140 companies. After analyzing their respective data sets, each finds a correlation coefficient of $r = 0.601$.

a) Who will have a smaller p-value?
   
   A. Bart
   B. Lisa
   C. Both will find the same p-value
   D. More information is needed to answer this question
b) Explain the reasoning your choice in part a).

**Height and finger length**

10.2.4 Recall from Exercise 10.1.13 that the data file **HeightAnd FingerLength** contains data on the height (inches) and the length of index finger of dominant hand (cm) for students in a statistics class. State in words the appropriate null and alternative hypotheses to test whether there is an association between height and length of index finger.

10.2.5 Refer to the previous exercise about the association between height and length of index finger. Recall that the data file **HeightAnd FingerLength** contains data on the height (inches) and the length of index finger of dominant hand (cm) for students in a statistics class. Describe how one might use everyday items (for example, coins, dice, cards, etc.) to conduct a tactile simulation-based test of the hypotheses. Be sure to clearly describe how the p-value will be computed from the simulation.

10.2.6 Refer to the previous exercise about the association between height and length of index finger. Recall that the data file **HeightAnd FingerLength** contains data on the height (inches) and the length of index finger of dominant hand (cm) for students in a statistics class.

The correlation coefficient for the sample data is 0.474. A simulated null distribution is shown to test whether the sample data provide evidence of a linear association between height and length of index finger.

![Graph showing a simulated null distribution with mean 0.004, SE 0.172, and num samples 1000.](image)
On the null distribution indicate the region that denotes the p-value, being sure to include any number that you are using to decide on the region. Also, make sure that you answer is consistent with the hypotheses chosen in part a).

**Used Honda Civics***

10.2.7 The data in the file *UsedHondaCivics* come from a sample of used Honda Civics listed for sale online in July 2006. The variables recorded in this data file are the car’s age (calculated as 2006 minus year of manufacture) and price. Consider conducting a simulation analysis to test whether the sample data provide strong evidence of a linear association between a car’s price and age in the population.

a) Identify the type of study: randomized experiment or observational study.

b) Identify the experimental/observational units in this study.

c) Identify the two variables of interest, and whether each is categorical or quantitative. Which variable do you think makes more sense to use as the explanatory variable, and which as the response variable?

d) State the appropriate null hypothesis in words.

e) State the relevant alternative hypothesis in words.

f) Enter the data into the *Corr/Regression* applet. Produce a scatterplot, and comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)

g) Report the value of the correlation coefficient. Does the value of the correlation coefficient support your answer in part f)? Explain how you are deciding.

h) Describe how one might use everyday items (for example, coins, dice, cards, etc.) to conduct a tactile simulation-based test of the hypotheses. Be sure to clearly describe how the p-value will be computed from the simulation.

10.2.8 Refer to the previous exercise about the association between the age and price of used Honda Civics. Consider conducting a test of significance to investigate whether the sample data provide strong evidence of an association between a car’s price and age in the population of used Honda Civics. The data are in the file *UsedHondaCivics*.
a) Enter the data into the Corr/Regression applet, and conduct a simulation-based test of significance using the correlation coefficient as the statistic. Report the approximate p-value.

b) Interpret the p-value reported in (a).

c) Summarize your conclusion (significance, causation, generalization) from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

House prices

10.2.9 Recall from Exercise 10.1.1 that the data file HousePrices contains data on prices ($) and sizes (in square feet) for a random sample of houses that sold in the year 2006 in Arroyo Grande, California.

a) State in words the appropriate null and alternative hypotheses to test whether there is a linear association between prices and sizes of houses.

b) Describe how one might use everyday items (for example, coins, dice, cards, etc.) to conduct a tactile simulation-based test of the hypotheses. Be sure to clearly describe how the p-value will be computed from the simulation.

10.2.10 Refer to the previous exercise about the association between prices and sizes of houses. Recall that the data file HousePrices contains data on prices ($) and sizes (in square feet) for a random sample of houses that sold in the year 2006 in Arroyo Grande, California.

a) Enter the data into the Corr/Regression applet, and conduct a simulation-based test of significance. Report the approximate p-value.

b) Interpret the p-value reported in part (a).

c) Summarize your conclusion (significance, causation, generalization) from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.
Roller coasters*

10.2.11 Recall from Exercise 10.1.12 that the data file RollerCoasters contains data (collected from www.rcdb.com) on the maximum speed (miles per hour) and the maximum height (feet) for each of a sample of roller coasters in the U.S.

a) State in words the appropriate null and alternative hypotheses to test whether is a linear association between maximum speed and maximum height of such roller coasters.

b) Describe how one might use everyday items (for example, coins, dice, cards, etc.) to conduct a tactile simulation-based test of the hypotheses. Be sure to clearly describe how the p-value will be computed from the simulation.

10.2.12 Refer to the previous exercise about the association between maximum speed and maximum height of roller coasters. Recall that the data file Rollercoasters contains data on the maximum speed (miles per hour) and the maximum height (feet) for each of a sample of roller coasters in the U.S.

a) Enter the data into the Corr/Regression applet, and conduct a simulation-based test of significance. Report the approximate p-value.

b) Interpret the p-value reported in part (a).

c) Summarize your conclusion (significance, causation, generalization) from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

Relationship between parents’ and their children’s heights

10.2.13 Recall from Exercise 10.1.16 that the data file MomandChildHeights contains data for students in a statistics class on the following two variables: student’s height (inches) and biological mother’s height (inches).

a) State in words the appropriate null and alternative hypotheses to test whether is a linear association between a person’s height and their biological mother’s height.
b) Describe how one might use everyday items (for example, coins, dice, cards, etc.) to conduct a tactile simulation-based test of the hypotheses. Be sure to clearly describe how the p-value will be computed from the simulation.

10.2.14 Refer to the previous exercise about the association between a person’s height and their biological mother’s height. Recall that the data file **MomandChildHeights** contains data for students in a statistics class on the following two variables: student’s height (inches) and biological mother’s height (inches).

a) Enter the data into the **Corr/Regression** applet, and conduct a simulation-based test of significance. Report the approximate p-value.
b) Interpret the p-value reported in part (a).
c) Summarize your conclusion (significance, causation, generalization) from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

10.2.15 Refer back to the previous exercises that explores the association between children’s heights and (biological) mothers’ heights. The data file **DadandChildHeights** contains data for students in a statistics class on the following two variables: student’s height (inches) and biological father’s height (inches).

a) State in words the appropriate null and alternative hypotheses to test whether is a linear association between a person’s height and their biological father’s height.
b) Describe how one might use everyday items (for example, coins, dice, cards, etc.) to conduct a tactile simulation-based test of the hypotheses. Be sure to clearly describe how the p-value will be computed from the simulation.

10.2.16 Refer to the previous exercise about the association between a person’s height and their biological father’s height. Recall that the data file **DadandChildHeights** contains data for students in a statistics class on the following two variables: student’s height (inches) and biological father’s height (inches).
Chapter 10: Two Quantitative Variables

a) Enter the data into the Corr/Regression applet, and conduct a simulation-based test of significance. Report the approximate p-value.
b) Interpret the p-value reported in part (a).
c) Summarize your conclusion (significance, causation, generalization) from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

**Scrabble names**

10.2.17 Recall from Exercise 10.1.19 that in a statistics class, students were asked to calculate the number of Scrabble points their names would earn. For example, the Scrabble score for the name Tom Sawyer would be 17 points. The data file ScrabbleNames contains data on the following two variables: number of letters in a student’s name and the corresponding Scrabble score.

a) State in words the appropriate null and alternative hypotheses to test whether there is a linear association between the number of letters in a person’s name and the Scrabble points their name earns.
b) Describe how one might use everyday items (for example, coins, dice, cards, etc.) to conduct a tactile simulation-based test of the hypotheses. Be sure to clearly describe how the p-value will be computed from the simulation.

10.2.18 Refer back to the previous exercise about exploring the relationship between the number of letters in a name and the corresponding Scrabble score. Recall that the data file ScrabbleNames contains data on the following two variables: number of letters in a student’s name and the corresponding Scrabble score.

a) Enter the data into the Corr/Regression applet, and conduct a simulation-based test of significance. Report the approximate p-value.
b) Interpret the p-value reported in part (a)
c) Summarize your conclusion (significance, causation, generalization) from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

10.2.19 Refer back to the previous exercise about exploring the relationship between the number of letters in a name and the corresponding Scrabble score. Students were also asked to calculate the points per letter, that is, \( \text{ratio} = \frac{\text{Scrabble score}}{\text{number of letters in name}} \). The data file ScrabbleRatio contains Scrabble points for a student’s name and the corresponding ratio.
a) State in words the appropriate null and alternative hypotheses to test whether is a linear association between the Scrabble points for a student’s name and the corresponding ratio.

b) Describe how one might use everyday items (for example, coins, dice, cards, etc.) to conduct a tactile simulation-based test of the hypotheses. Be sure to clearly describe how the p-value will be computed from the simulation.

10.2.20 Refer back to the previous exercise. Recall that the data file ScrabbleRatio contains Scrabble points for a student’s name and the corresponding ratio of Scrabble points to number of letters.

a) Enter the data into the Corr/Regression applet, and conduct a simulation-based test of significance. Report the approximate p-value.

b) Interpret the p-value reported in part (a).

c) Summarize your conclusion (significance, causation, generalization) from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

TV and life expectancy

10.2.21 Recall from Exercise 10.1.22 that the data file TVLife provides information on life expectancy and number of televisions per thousand people in a sample of 22 countries, as reported by the 2006 World Almanac and Book of Facts.

a) State in words the appropriate null and alternative hypotheses to test whether is a linear association between the life expectancy and number of televisions per thousand people in such countries.

b) Describe how one might use everyday items (for example, coins, dice, cards, etc.) to conduct a tactile simulation-based test of the hypotheses. Be sure to clearly describe how the p-value will be computed from the simulation.

10.2.22 Refer back to the previous exercise. Recall that the data file TVLife provides information on life expectancy and number of televisions per thousand people in a sample of 22 countries, as reported by the 2006 World Almanac and Book of Facts.

a) Enter the data into the Corr/Regression applet, and conduct a simulation-based test of significance. Report the approximate p-value.
b) Interpret the p-value reported in part (a).
c) Summarize your conclusion (significance, causation, generalization) from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

Section 10.3

10.3.1* Which of the following does the method of least squares minimize?
   A. Sum of vertical distances between observations and the line
   B. Sum of squared vertical distances between observations and the line
   C. Sum of perpendicular distances between observations and the line
   D. Sum of squared perpendicular distances between observations and the line
   E. Sum of horizontal distances between observations and the line
   F. Sum of squared horizontal distances between observations and the line

10.3.2 Explain what is wrong with the following statements about correlation and regression.
   a) The correlation between the number of years of education and yearly income is 1.23.
   b) The regression equation that describes the relationship between the age of a used Ford Mustang and its value is \( \hat{y} = -1000x + 10000 \) where the explanatory variable is the age of the automobile (in years) and the response is its value (in dollars). The correlation describing this relationship is 0.83.

10.3.3* It can be shown that the sum of residuals from a least squares line always equals zero.
   a) Does it follow from this result that the mean of the residuals from a least squares line always equals zero? Explain briefly.
   b) Does it follow from this result that the median of the residuals from a least squares line always equals zero? Explain briefly.

10.3.4 Answer the following questions about slope coefficients.
   a) Can a slope coefficient ever be negative?
   b) Can a slope coefficient ever be greater than one?
   c) Can a slope coefficient ever equal zero?
10.3.5* Reconsider the previous exercise. For each of the descriptions presented that is indeed possible, produce a scatterplot of hypothetical data to show that the description is indeed possible.

10.3.6 Reconsider the previous two exercises. Answer the questions with regard to the intercept coefficient rather than the slope coefficient.

Legos*

10.3.7 A statistics professor went to the lego.com website in February 2014 and recorded the number of pieces and the sales price for 157 Lego products listed there. The data appear in the datafile legos.

a) Which variable do you think makes more sense to use as the explanatory variable, and which as the response variable?

b) Enter the data into the Corr/Regression applet. Produce a scatterplot, and comment on the association between the variables as revealed in the scatterplot. (Remember to comment on form, direction, strength, and unusual observations.)

c) Report the value of the correlation coefficient.

d) Calculate the value of $r^2$, and interpret what the value means in this context.

10.3.8 Reconsider the previous exercise and the datafile legos.

a) Determine the equation of the least squares line for predicting price from number of pieces. Report this equation using good statistical notation.

b) Interpret the value of the slope coefficient in this context.

c) Interpret the value of the intercept coefficient. Is this a context for which the value of the intercept provides relevant information? Explain.

10.3.9 Reconsider the previous two exercises and the datafile legos. The equation of the least squares line turns out to be $\text{price} = 4.86 + 0.105 \times \text{pieces}$, and the value of $r^2 = 0.949$. Identify which of the following interpretations are correct, which are incorrect, and which are correct but incomplete or poorly worded.

a) The Legos cost 10.5 cents per piece.

b) For each additional dollar in price, the predicted number of pieces in the set increases by about 10.5.

c) For each additional piece in the set, the predicted price increases by about 10.5 cents.
d) The price goes up 10.5 cents for each additional piece in the set.
e) The predicted price for a set with 0 pieces is $4.86.
f) The predicted price increases by $10.5 for each additional 100 pieces in the set.

10.3.10 Reconsider the previous three exercises and the datafile legos. The equation of the least squares line turns out to be \( \text{price} = 4.86 + 0.105 \times \text{pieces} \), and the value of \( r^2 = 0.949 \). Identify which of the following interpretations are correct, which are incorrect, and which are correct but incomplete or poorly worded.

a) 10.5% of the observed variability in prices is explained by the least squares line with number of pieces.
b) 94.9% of the observed variability in prices is explained by the least squares line with number of pieces.
c) 94.9% of the prices fall on the least squares line based on number of pieces.
d) The least squares line correctly predicts the price for 94.9% of the Lego products.

10.3.11 Reconsider the previous four exercises and the datafile legos.

a) Determine the predicted price for a product with 500 pieces.
b) Determine the predicted price for a product with 1500 pieces.
c) Determine the difference between the two predictions in parts a) and b). Also indicate how you could have determined this value directly from the equation of the least squares line.
d) Determine the predicted price for a product with 5000 pieces.
e) Do you feel very confident with this prediction in part e)? Explain why or why not.

10.3.12 Reconsider the previous five exercises and the datafile legos. The last product listed in the datafile has 415 pieces and a price of $49.99.

a) Determine the predicted price for such a product.
b) Determine the residual value for this product.
c) Interpret what this residual value means.
d) Does the product fall above or below the least squares line in the graph? Explain how you can tell, based on its residual value.

10.3.13 Reconsider the previous six exercises and the datafile legos. This is very unrealistic, but suppose that one of the products were to be offered at a price of $0.
a) Would you expect this change to affect the least squares line very much? Explain.
b) For which one product would you expect this change to have the greatest impact on the least squares line? Explain how you choose this product.
c) Change the price to $0 for the product that you identified in part b). Report the (new) equation of the least squares line and the (new) value of $r^2$. Have these values changed considerably?

Crickets

10.3.14 Consider the following two scatterplots based on data gathered in a study of 30 crickets, with temperature measured in degrees Fahrenheit and chirp frequency measured in chirps per minute:

a) If the goal is to predict temperature based on a cricket’s chirps per minute, which is the appropriate scatterplot to examine— the one on the left or the one on the right? Explain briefly.

One of the following is the correct equation of the least squares line for predicting temperature from chirps per minute:

A. predicted temperature = 35.78 + 0.25 chirps per minute
B. predicted temperature = -131.23 + 3.81 chirps per minute
C. predicted temperature = 83.54 – 0.25 chirps per minute

b) Which is the correct equation? Circle your answer, and explain briefly.
c) Use the correct equation to predict the temperature when the cricket is chirping at 100 chirps per minute.
d) Interpret the value of the slope coefficient, in this context, for whichever equation you think is the correct one.
Cat jumping*

10.3.15 Harris and Steudel (2002) studied factors that might be associated with the jumping performance of domestic cats. They studied 18 cats, using takeoff velocity (in centimeters per second) as the response variable. They used body mass (in grams), hind limb length (in centimeters), muscle mass (in grams), and percent body fat, in addition to sex, as potential explanatory variables. The data can be found in the CatJumping. A scatterplot of takeoff velocity vs. body mass is shown here:

```
a) Describe the association between these variables.
b) Use the Corr/Regression applet to determine the equation of the least squares line for predicting a cat’s takeoff velocity from its mass.
c) Interpret the value of the slope coefficient in this context.
d) Interpret the value of the intercept coefficient. Is this a context in which the intercept coefficient is meaningful?
e) Determine the proportion of variability in takeoff velocity that is explained by the least squares line with mass.
```

10.3.16 Reconsider the previous exercise and the CatJumping datafile.

a) Determine the predicted takeoff velocity for a cat with a mass of 5000 grams (which is about 11 pounds).
b) Determine the predicted takeoff velocity for a cat with a mass of 10,000 grams. Also explain why it’s not advisable to have much confidence in this prediction.
c) Determine the predicted takeoff velocity and the residual value for the cat with the largest mass. Also interpret what this residual value means.

10.3.17 Reconsider the previous two exercises. Answer the following based on the scatterplot presented above. Do not bother to perform any calculations.
   a) Which cat has the largest predicted value for its takeoff velocity?
   b) Which cat has the smallest predicted value for its takeoff velocity?
   c) Which cat has the largest residual value?
   d) Which cat has the smallest residual value?

10.3.18 Reconsider the previous three exercises and the CatJumping datafile. Investigate the association between the response variable (takeoff velocity) and the other explanatory variables (hind limb length, muscle mass, percent body fat).
   a) Select the explanatory variable that has the strongest association with the response. Describe this association.
   b) Report the equation of the least squares line for predicting the cat’s takeoff velocity using this explanatory variable.
   c) Interpret the value of the slope coefficient.
   d) Determine and interpret the value of $r^2$.

Honda civic pricing
10.3.19 The data in the file UsedHondaCivics come from a sample of used Honda Civics listed for sale online in July 2006. The variables recorded are the car’s year of manufacture, age (calculated as 2006 minus year of manufacture), mileage, and price.
   a) Identify the observational units.
   b) Produce a scatterplot of price vs. age. Describe the association revealed in the graph.
   c) Determine the least squares line for predicting price from age, and produce a scatterplot with the least squares line superimposed.
   d) Report and interpret the value of the slope coefficient.
   e) What percentage of the variability in car prices is explained by knowing the car’s age?
Textbook prices*

10.3.20 Two Cal Poly freshmen gathered data on a random sample of textbooks from the campus bookstore in November of 2006. Two of the variables recorded were the price of the book and the number of pages that it contained. These data are in the file TextbookPrices.

a) Identify the explanatory and response variables in this study.

b) Determine the equation of the least squares line for predicting price from number of pages. Report the equation, being sure to use good statistical notation.

c) Use the least squares line to predict the price of a 500-page textbook. Then do the same for a 1500-page textbook. Which prediction would you have more confidence in? Explain.

d) Interpret what the slope coefficient means in this context.

e) Determine the proportion of variability in textbook prices that is explained by knowing the number of pages in the book.

Day hikes

10.3.21 The book Day Hikes in San Luis Obispo County lists information about 72 hikes, including the distance of the hike (in miles), the elevation gain of the hike (in feet), and the time that the hike is expected to take (in minutes). Consider the scatterplot below, with least squares regression lines superimposed:

a) Report the value of the slope coefficient for predicting time from distance.

b) Write a sentence interpreting the value of the slope coefficient for predicting time from distance.
c) Use the line to predict how long a 4-mile hike will take.

d) Would you feel more comfortable using the line predict the time for a 4-mile hike or for a 12-mile hike? Explain your choice.

e) The value of the correlation coefficient between time and distance is 0.916, and the value of \( r^2 = 0.839 \). Complete this sentence to interpret what this value means:

\[
83.9\% \text{ of } \underline{\text{ }} \text{ is explained by } \underline{\text{ }}.
\]

10.3.22 Reconsider the previous exercise. The following scatterplot displays hiking time vs. elevation gain, with the least squares line superimposed:

a) Report the value of the slope coefficient for predicting time from elevation gain.

b) Write a sentence interpreting the value of the slope coefficient for predicting time from elevation gain.

c) Use the line to predict how long a hike with an 800-feet elevation gain will take.


d) Would you feel more comfortable using the line predict the time for a hike with an 800-feet elevation gain or a hike with a 2800-feet elevation gain? Explain your choice.

e) The value of the correlation coefficient between time and distance is 0.344, and the value of \( r^2 = .119 \). Complete this sentence to interpret what this value means:

\[
11.9\% \text{ of } \underline{\text{ }} \text{ is explained by } \underline{\text{ }}.
\]

10.3.23 The slope \( b \) and intercept \( a \) of a least squares line can be calculated from the means and standard deviations of the two variables, along with the correlation coefficient, as follows:

\[
b = r \frac{s_y}{s_x} \quad \text{and} \quad a = \bar{y} - b \bar{x}.
\]

These summary statistics for the day hikes described in the previous two exercises are:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Correlation with time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td>3.283</td>
<td>1.856</td>
<td>0.916</td>
</tr>
<tr>
<td>Elevation gain</td>
<td>333.2</td>
<td>355.5</td>
<td>0.344</td>
</tr>
</tbody>
</table>
a) Use these statistics to determine the slope $b$ and intercept $a$ for predicting time from distance.
b) Use these statistics to determine the slope $b$ and intercept $a$ for predicting time from elevation gain.
c) Use the least squares line to predict the time needed for a hike with distance equal to the mean of the distances.
d) Use the least squares line to predict the time needed for a hike with elevation gain equal to the mean of the elevation gains.
e) What do you notice about your predictions in parts c) and d)?

**Introductory Statistics Tests**

**10.3.24** The following scatterplot represents scores on test 2 and test 3 in Introductory Statistics for a random sample of students.

a) In the figure above, there appears to be a positive relationship between scores on test 2 and test 3. Does this mean that a person scores above the mean on test 2 is expected to also score above the mean on test 3? What about those who score below the mean on test 2, are they expected to score below the mean on test 3?

The regression equation for these data is $\hat{y} = 552 + 0.3708x$, where $x$ is the score on test 2 and $\hat{y}$ is the expected score on test 3. The mean for test 2 was 79.8 and the mean for test 3 was 84.8.
b) If a student scores a 79.8 on test 2, what is their predicted score for test 3? If that student then actually gets a 90 on test 3, what is the residual?

c) If someone scores 10 points above the mean on test 2, how many points above the mean on test 3 is their expected score?

d) If someone scores 10 points below the mean on test 2, how many points below the mean on test 3 is their expected score?

10.3.25 Reconsider the previous exercise.

a) Determine the predicted score on test 3 for a student who scores 55 on test 2.

b) Determine the predicted score on test 3 for a student who scores 95 on test 2.

c) Of the two students in parts a) and b), which student is predicted to achieve a higher score on test 3 than he/she scored on test 2? Which student is predicted to achieve a lower score on test 3 than he/she scored on test 2? (This phenomenon is known as the regression effect.)

Section 10.4

10.4.1* When testing the hypothesis that there is no association (null) vs. an association (alternative) you can use either the sample correlation coefficient or the sample slope as the statistic. How will the p-values compare when using both approaches on the same data set?

A. The result from using slope as our statistic is equivalent to using the correlation coefficient as our statistic in a test; therefore our p-values will be identical.

B. The result from using slope as our statistic is similar to using the correlation coefficient as our statistic in a test; therefore our p-values will be similar.

C. As slope and the correlation coefficient measure two different things, our p-values will, most likely, not be similar.

D. There is no way to tell without running the tests.

Foot length and height

10.4.2 Suppose a sample of 10 people had their foot lengths measured (in cm) and heights (in inches) and the results, along with the regression line are shown in the following scatterplot. The equation of the regression line, correlation coefficient and $r^2$ are also shown.
Which one of the following applet screen shots is the appropriate way to test to determine if there was a positive relationship between people’s foot length and height?

A

B

C

D

Count Samples | Greater Than | 0.7 | Count
Count = 88/1000 (0.0880)

Count Samples | Greater Than | 1.10 | Count
Count = 11/1000 (0.0110)

Count Samples | Beyond | 0.7 | Count
Count = 139/1000 (0.1390)

Count Samples | Beyond | 1.10 | Count
Count = 25/1000 (0.0250)
10.4.3* Referring to the previous exercise, describe how you would construct a null distribution for this situation by hand using the slope of the regression line as your statistic. Assume you have 20 slips of paper.

10.4.4 Using foot length and height data, we shuffled the y-values 100 times and each time plotted the resulting regression line in the graph below. The regression line for the observed data is the one being pointed at with the dashed arrow. Based on the plot of these lines, what can you say about the p-value if we are testing to see whether there is a positive association between height and foot length? Explain.

**Height and leap**

10.4.5 Suppose you are testing to see whether there is an association between a person’s height and their vertical leap. You’ve collected data on this from 20 people. Describe how you would construct a null distribution for this situation by hand using the slope of the regression line as your statistic. Assume you have 20 slips of paper containing the 20 heights and 20 slips of paper containing the 20 vertical leap distances.

**Textbook prices**

10.4.6 A random sample of 15 textbooks in a campus bookstore was selected. Two of the variables recorded were the price of the book (in dollars) and the number of pages it contained. The data are displayed in the following scatterplot.
a) Is this an experiment or an observational study?
b) What are the observational units?
c) The equation of the least squares regression line for predicting price from number of pages is: \( \text{price} = 14.11 + 0.13(\text{pages}) \)
   i. Use the regression equation to predict the price of a 500-page textbook.
   ii. Use the regression equation to predict the price of a 1500-page textbook.
   iii. Which prediction, the one for the 500 page book or 1500 page book, should you have more confidence in? Explain.

10.4.7 Refer to the information in the previous exercise to answer the following.
   a) State the null and alternative hypotheses for a test of possible association between pages and price.
   b) The following null distribution was created to test the hypotheses stated in the previous question using slope as the statistic.
i. Based on information shown in the null distribution, how many standard deviations is our observed statistic above the mean of the null distribution? (i.e. What is the standardized statistic?)

ii. Based on your standardized statistic, do you have strong evidence of an association between number of pages and price of textbooks? Explain.

10.4.8 Reconsider the previous two exercises about textbook prices. The equation of the least squares regression line for predicting price from number of pages is:

\[ \text{Price} = 14.11 + 0.13(\text{pages}) \]

a) Interpret what the slope coefficient means in the context of pages and price.

b) Interpret the intercept. Is this an example of extrapolation? Why or why not?

Sleep and maze performance*

10.4.9 Student researchers asked their subjects how much sleep they had the previous night (in hours) and then timed how long it took them (in seconds) to complete a paper and pencil maze. The results are shown in the scatterplot along with the regression line.

The following null distribution was created to test the association between number of time and the amount of sleep using slope as the statistic. The null and alternative hypotheses for this test can be written as: Null: No
association between sleep and time in the population, Alt: Association between sleep and time in the population.

- **i.** Based on information shown in the null distribution, how many standard deviations is our observed statistic above the mean of the null distribution? (i.e. What is the standardized statistic?)

- **ii.** Based on your standardized statistic, do you have strong evidence of an association between the time it takes someone to complete the maze and how much sleep they got the night before? Explain.

10.4.10 Reconsider the previous exercise about the amount of sleep (in hours) obtained in the previous night and time to complete a paper and pencil maze (in seconds). The equation of the least squares regression line for predicting price from number of pages is:

\[ \text{time} = 190.33 - 7.36\text{sleep}. \]

- **a)** Interpret what the slope coefficient means in the context of sleep and time to complete the maze.
- **b)** Interpret the intercept. Is this an example of extrapolation? Why or why not?

**Weight loss and protein**

10.4.11 In a research study was looking to see if there was an association between weight loss and the amount of a certain protein in a person’s body fat, the researchers measured a number of different attributes in their 39 subjects at the beginning of the study. The article reported, “These subjects were clinically and ethnically heterogeneous.” Two of the variables they measured were
body mass index (BMI) and total cholesterol. The results are shown in the scatterplot along with the regression line.

![Scatterplot of BMI vs Cholesterol](image)

**a)** What are the observational units in the study?

**b)** The equation of the least squares regression line for predicting total cholesterol from BMI is: \( \text{Cholesterol} = 162.50 - 0.9058 \times \text{BMI} \). The following null distribution was created to test the association between people’s total cholesterol number and their BMI using slope as the statistic. The null and alternative hypotheses for this test can be written as: Null: No association between cholesterol and BMI in the population, Alt: Association between cholesterol and BMI in the population.

**i.** Based on information shown in the null distribution, how many standard deviations is our observed statistic below the mean of the null distribution? (i.e. What is the standardized statistic?)

**ii.** Based on your standardized statistic, do you have strong evidence of an association between a people’s total cholesterol and their BMI? Explain.
10.4.12 Reconsider the previous exercise about the cholesterol and BMI. The equation of the least squares regression line obtained was
\[ \text{cholesterol} = 162.56 - 0.9658(\text{BMI}) \].

a) Interpret what the slope coefficient means in the context of cholesterol and BMI.
b) Interpret the intercept. Is this an example of extrapolation? Why or why not?
c) Based on your confidence interval do you have strong evidence of an association between cholesterol and BMI? Explain.

Honda Civic prices*

10.4.13 The data in the file \texttt{UsedHondaCivics} come from a sample of used Honda Civics listed for sale online in July 2006. The variables recorded are the car’s year of manufacture, age (calculated as 2006 minus year of manufacture) and price. Consider conducting a simulation analysis to test whether the sample data provide strong evidence of an association between a car’s price and age in the population, in terms of the population slope.

a) State the appropriate null and alternative hypotheses.
b) Conduct a simulation analysis with 1000 repetitions. Describe how to find your p-value from your simulation results, and report this p-value.
c) Summarize your conclusion from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

10.4.14 Reconsider the previous exercise on prices of Honda Civics.

a) Find the regression equation that predicts the price of the car given its age.
b) Interpret the slope and intercept of the regression line.

Weight and haircut price

10.4.15 In a survey of statistics students at Hope College, two of the questions asked were their weight (in pounds) and the cost of their last haircut, including any hair treatments (in dollars). The data can be found in the file \texttt{WeightHaircut}. Let’s explore whether or not there is evidence of a strong association between weight and cost of haircuts.

a) State the appropriate null and alternative hypotheses.
b) Conduct a simulation analysis with 1000 repetitions. What is your p-value?
c) Summarize your conclusion from this simulation analysis. If there is an association, describe if it is positive or negative in the context of the study.

d) Can you conclude a cause and effect between weight and cost of a haircut? Why or why not?

**10.4.16** Reconsider the previous exercise on weight and haircut price.

a) Find the regression equation that predicts the haircut price based on weight.

b) Interpret the slope and intercept of the regression line.

**Height and haircut price**

**10.4.17** In a survey of statistics students at Hope College, two of the questions asked were their height (in inches) and the cost of their last haircut, including any hair treatments (in dollars). The data can be found in the file **HeightHaircut**. Let’s explore whether or not there is evidence of a strong association between height and cost of haircuts.

a) State the appropriate null and alternative hypotheses.

b) Conduct a simulation analysis with 1000 repetitions. What is your p-value?

c) Summarize your conclusion from this simulation analysis. If there is an association, describe if it is positive or negative in the context of the study.

d) Can you conclude a cause and effect between height and cost of a haircut? Why or why not?

**10.4.18** Reconsider the previous exercise on height and haircut price.

a) Find the regression equation that predicts the haircut price based on height.

b) Interpret the slope and intercept of the regression line.

**Age and BMI**

**10.4.19** Researchers in a clinical study collected information from their subjects at the beginning of the study. Two of the variables were body mass index (BMI) and age. We are interested in seeing whether there is an association between BMI and age. The data from the study can be found in the file **AgeBMI**.

a) State the appropriate null and alternative hypotheses.

b) Conduct a simulation analysis with 1000 repetitions. What is your p-value?
c) Summarize your conclusion from this simulation analysis. If there is an association, describe if it is positive or negative in the context of the study.

10.4.20 Reconsider the previous exercise on age and BMI. We are interested in seeing whether there is an association between BMI and age. The data from the study can be found in the file AgeBMI.

a) Conduct a new simulation analysis with 1000 repetitions using slope as the statistic. What is your p-value?

b) Using the same null distribution as in part (a), use correlation as the statistic. What is the p-value?

c) Are you two p-values exactly the same? (The two p-values should be the same; they may be just slightly different due to rounding of the statistics.)

Missing class and GPA*

10.4.21 In a survey of statistics students at Hope College, two of the questions asked were their current grade point average (GPA) and how many classes they failed to attend during the past three weeks at the college. The data can be found in the file MissClassGPA. Let’s explore whether or not there is evidence of a strong association between these two variables.

a) State the appropriate null and alternative hypotheses

b) Conduct a simulation analysis with 1000 repetitions. What is your p-value?

c) Summarize your conclusion from this simulation analysis. If there is an association, describe if it is positive or negative in the context of the study.

d) Can you conclude a cause and effect between missing class and GPA? Why or why not?

Sleep and GPA

10.4.22 In a survey of Hope College students, two of the questions asked were their current grade point average (GPA) and how much sleep the got (in hours) on the previous school night. The data can be found in the file SleepGPA. Let’s explore whether or not there is evidence of a strong association between these two variables.

a) What is the value of the correlation? Based on this number, does it appear there is a strong association between our two variables?

b) What is the value of the slope of the regression line? What does this number mean in terms of sleep and GPA?
c) State the appropriate null and alternative hypotheses.
d) Conduct a simulation analysis with 1000 repetitions. What is your p-value?
e) Based on your p-value, is there a strong association between sleep and GPA for Hope College students?
f) There should be some conflict between your estimate of the strength of association in part (a) and your answer to part (e). Why did that occur with this data set?

**Stroop effect**

10.4.23 You may have seen the test or puzzle where names of colors are written out in colors not denoted by the name. For example the word GREEN might be written with red ink. The job of the person completing the test is to say the word and not the color. This interference in the reaction time of saying the words is called the Stroop effect. Student researchers at Hope College wanted to see if there was an association between age and the time it takes people to read a list of 20 of these colored words. Their results can be found in the data file *StroopAgeTime*. The age is given in years and the time to complete the task is given in seconds. Put the data in the Corr/Regression applet and complete the following.

a) Write out the null and alternative hypotheses for this study using a two-sided alternative.
b) What is the slope of the regression line? What does that number mean in the context of age and time to complete the task?
c) Conduct at least 1000 shuffles of the data, find and report the resulting p-value.
d) Summarize your conclusion from this analysis.

10.4.24 Reconsider the previous exercise on the Stroop effect.
a) There was one point in the scatterplot that was a clear outlier. One of the researchers’ great aunt, who was 87 years old, took 100.6 seconds to complete the task. This was almost three times as long as the next longest time. Remove that point from the data set and find a new p-value. Are your results significant?
b) After the great aunt’s point is removed from the scatterplot, there now appears as a couple of other points representing an 8 year-old and a 10 year-old might be outliers. Perhaps an association between age and time won’t be demonstrated when the subjects are too young. For this reason, delete all the points representing anyone 10 and under. There should be 5 of these. Find a new p-value. Are your results significant?
c) With the 87 year-old included, there is a significant association. When the result from the 87 year-old is deleted from the data, there is not a significant association. When results from everyone 10 and under are deleted from the study, there is a significant association. So, how would you answer the question, is there an association between age and time needed to complete this task of reading of color names? Explain.

Section 10.5

10.5.1* Which one of the following is a validity condition for theory-based tests for regression?
A. The variability of the points around the regression line should not differ as you slide along the $x$-axis.
B. There should be at least 10 success and 10 failures in the data.
C. The variability of the explanatory variable should be the same as the variability of the response variable.
D. The scatterplot should show a positive association between the explanatory and response variables.

10.5.2 For a given dataset, a test of association based on a slope is equivalent to a test of association based on a correlation coefficient. Being equivalent means which of the following is true?
A. The confidence intervals for the population correlation and population slope will be the same.
B. The observed correlation will be the same as the observed slope of the regression line.
C. The p-value will be the same whether you use correlation as the statistic or the slope of the regression line as the statistic.
D. All the above.

10.5.3* The theory-based p-value given in the regression table in the Corr/Regression applet is:
A. A two-sided p-value.
B. A one-sided p-value.
C. Either a one-sided or two-sided p-value depending how count your samples (greater than, less than, or beyond) in the simulation part of the applet.
D. A one-sided p-value when you using correlation as the statistic, but a two-sided p-value when you are using slope as the statistic.

10.5.4 To explore the relationship between the runtime and profit made by movies, data were collected on a random sample of 15 movies released in the last 4 years. The p-value corresponding to the two-sided test for the slope turned out to be 0.06.

a) Based on the p-value, is it okay to conclude that “there is evidence that there is no association between runtime and profit made by movies”? How are you deciding?

b) Suppose that you were to create a 90% confidence interval for the population slope, using the data from this study. Would this confidence interval contain 0? How are you deciding?

c) How, if at all, would the p-value change if you wanted to test whether there is a positive association between runtime and profit made by movies? Explain.

10.5.5* If there are influential points or outliers in your data, why is it a good idea to run the analysis with those points removed?

10.5.6 For the data in each of the following scatterplots, are the validity conditions met to run a theory-based test for regression? If the answer is no on any of them, explain why not.

a)  

b) 

c)  
d)
Sleep and maze completion*

10.5.7 A regression table is shown based on data used to test an association between the amount of sleep someone had the previous night (in hours) and the time needed to complete a paper and pencil maze (in seconds). Sleep is the explanatory variable and time need to complete the maze is the response.

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>199.53</td>
<td>51.57</td>
<td>3.85</td>
<td>0.0003</td>
</tr>
<tr>
<td>Sleep</td>
<td>7.76</td>
<td>7.50</td>
<td>-1.03</td>
<td>0.3052</td>
</tr>
</tbody>
</table>

a) What is the regression equation where time to complete the maze is predicted from the amount of sleep?

b) If we were testing to the alternative hypothesis $\beta \neq 0$, what is the p-value?

c) If we were testing to the alternative hypothesis $\beta < 0$, what is the p-value?

BMI and glucose levels

10.5.8 Problems 10.5.8 and 10.5.9 refer to data on BMI and glucose levels.

In a research study that was looking to see if there was an association between weight loss and the amount of a certain protein in a person’s body fat, the researchers measured a number of different attributes in their 39 subjects at the beginning of the study. The article reported, “These subjects were clinically and ethnically heterogeneous.” Two of the variables they measured were body mass index (BMI) and blood glucose levels and the results are in the file BMI-Glucose. We are interested in seeing whether there is an association between BMI and glucose levels.
a) Describe the population slope in the context of the study, and assign a symbol to it.
b) State the appropriate null and alternative hypotheses in terms of the population slope, using the symbol used in (a).
c) Is a theory-based test appropriate? Why or why not?
d) Regardless of your answer to the previous question, use a theory-based approach to find the p-value to test the hypotheses stated in (b). Report and interpret this p-value, in the context of the study.
e) Summarize your conclusion, based on the p-value reported in (d).
f) If you were testing to determine if there was a positive association between BMI and glucose levels, what is the theory-based p-value?

10.5.9 Reconsider the previous exercise on BMI and blood glucose levels.

a) Determine a 95% confidence interval for the population slope coefficient, and interpret what this interval represents.
b) Based on your confidence interval, is there a positive association between BMI and blood glucose levels? Explain.

Height and Weight*

10.5.10 At the beginning of the semester, students in a statistics class were asked to give their estimates of how much their professor weighed (in pounds) and his height (in inches). We want to see if students who give large estimates for height will also give large estimates for weight. In other words, is there a positive association between these two variables. The data from the study can be found in the file HeightWeight. (The professor actually weighed about 160 pounds and was 72 inches tall.)

a) Describe the population slope in the context of the study, and assign a symbol to it.
b) State the appropriate null and alternative hypotheses in terms of the population slope, using the symbol used in (a).
c) Is a theory-based test appropriate? Why or why not?
d) Regardless of your answer to the previous question, use a theory-based approach to find the p-value to test the hypotheses stated in (b). Report and interpret this p-value, in the context of the study.
Age and BMI

10.5.11 Problems 10.5.11 and 10.5.12 refer to data on age and BMI. Researchers in a clinical study collected information from their subjects at the beginning of the study. Two of the variables were body mass index (BMI) and age. We are interested in seeing whether there is an association between these two variables. The data from the study can be found in the file AgeBMI.

a) Describe the population slope in the context of the study, and assign a symbol to it.
b) State the appropriate null and alternative hypotheses in terms of the population slope, using the symbol used in (a).
c) Is a theory-based test appropriate? Why or why not?
d) Use a theory-based approach to find p-value.
e) Summarize your conclusion, based on the p-value reported in (d).
f) If you were testing to determine if there was a positive association between age and BMI, what is theory-based p-value?

10.5.12 Reconsider the previous exercise on age and BMI.

a) Determine a 95% confidence interval for the population slope and interpret what this interval represents.
b) Based on your confidence interval, is there strong evidence for a positive association between age and BMI? Explain.
c) Based on a one-sided p-value, is there strong evidence for a positive association between age and BMI? Explain.
d) Your answers to parts (b) and (c) might contradict each other. Explain why that is.

Honda Civic prices

10.5.13 Problems 10.5.13 and 10.5.14 refer to data gathered on Honda Civic prices. The data in the file UsedHondaCivics come from a sample of used Honda Civics listed for sale online in July 2006. The variables recorded are the car’s year of manufacture, age (calculated as 2006 minus year of manufacture) and price. Consider conducting an analysis to test whether the sample data provide strong evidence of an association between a car’s price and age in the population, in terms of the population slope.

a) Describe the population slope in the context of the study, and assign a symbol to it.
b) State the appropriate null and alternative hypotheses in terms of the population slope, using the symbol used in (a).
c) Is a theory-based test appropriate? Why or why not?
d) Use a theory-based approach to find a p-value to test the hypotheses stated in (b).
e) Summarize your conclusion, based on the p-value reported in (d).

10.5.14 Reconsider the previous exercise on Used Honda Civics.

a) Determine a 95% confidence interval for the population slope coefficient, and interpret what this interval represents.
b) Based on your confidence interval, is there a negative association between age and price for used Honda Civics? Explain.

Textbook prices

10.5.15 Two Cal Poly freshmen gathered data on a random sample of textbooks from the campus bookstore in November of 2006. Two of the variables recorded were the price of the book and the number of pages that it contained. These data are in the file TextbookPrices.

a) Describe the population slope in the context of the study, and assign a symbol to it.
b) Suppose that you want to test whether there is a positive association between price and number of pages in the population of all textbooks for sale in the bookstore in November of 2006. State the appropriate null and alternative hypotheses in terms of the population slope, using the symbol used in (a).
c) Is a theory-based test appropriate? Why or why not?
d) Determine and report the value of the $t$ statistic and p-value for testing the hypotheses stated in (b).
e) Do the sample data provide strong evidence that number of pages is a significant predictor of price? Explain how you can tell.
f) Determine and interpret a 95% confidence interval for the value of the population slope.

Height and BMI*

10.5.16 Problems 10.5.16 and 10.5.17 refer to data gathered on height and BMI. The data in the file HeightBMI was obtained from a survey of statistics students. The variables recorded are the student’s heights (in inches) and their body mass index (BMI). Consider conducting an analysis to test whether the sample data provide strong evidence of an association between height and BMI, using slope as the statistic.

a) Describe the population slope in the context of the study, and assign a symbol to it.
b) State the appropriate null and alternative hypotheses in terms of the population slope, using the symbol used in (a).

c) Is a theory-based test appropriate? Why or why not?

d) Use a theory-based approach to find a p-value.

e) Summarize your conclusion, based on the p-value reported in (d).

10.5.17 Reconsider the previous exercise on height and BMI.

a) Determine a 95% confidence interval for the population slope coefficient, and interpret what this interval represents.

b) Based on your confidence interval, is there an association between height and BMI? Explain.

Gestation and Life Expectancy

10.5.18 The data in the file GestationLifeExpectancy gives the gestation period (in day) and the life expectancy (in years) for a sample of mammals. We will consider this sample to be representative of all mammals for these two variables. Consider conducting an analysis to test whether the sample data provide strong evidence of an association between gestation period and life expectancy.

a) What is the value of the slope of the regression line and what does it mean in terms of gestation period and life expectancy?

b) State the appropriate null and alternative hypotheses in terms of the population slope.

c) Use a theory-based approach to find a p-value. Do you have strong evidence of an association between the two variables?

d) Remember that a TYPE II error is not rejecting a false null hypothesis (or a missed opportunity). Do you think you made a TYPE II error for this exercise? Why or why not?

Sleep and height*

10.5.19 The data in the file SleepHeight.txt was obtained from a survey of statistics students. The variables recorded are the student’s heights (in inches) and the amount of sleep they typically get on a school night (in hours). Consider conducting an analysis to test whether the sample data provide strong evidence of a negative association between sleep and height.

a) Describe the population slope in the context of the study, and assign a symbol to it.
b) State the appropriate null and alternative hypotheses in terms of the population slope, using the symbol used in (a).

c) Use a theory-based approach to find a p-value. Do you have strong evidence of a negative association between sleep and height?

d) There are a couple of errors in the data set. It appears as though two students reported their heights as 55 inches and they sleep 10 hours a night. Because all of the answers on the survey were identical for these two entries, it would appear they came from the same person. It is also obvious to the instructor that there was nobody in class as short as 55 inches, so we know the height is an error. Eliminate these two points (they are the last ones on the list) and find a new p-value. Do you now have strong evidence of a negative association?

**GDP and infant mortality**

**10.5.20** The following scatterplot gives data from a random sample of 27 countries where the explanatory variable is infant mortality (deaths per 1000 live births) and the response is the gross domestic product (GDP) per capita (in thousands of dollars). The gross domestic product per capita is a representation of a country’s standard of living.

![Scatterplot of GDP and Infant Mortality](image)

a) Describe the relationship between infant mortality and GDP per capita. Is this relationship linear?

b) The regression line is also included in the scatterplot. Will this regression line be a good prediction of infant mortality given GDP per capita? Why or why not?
Sometimes we can transform data that don’t fit a linear pattern in such a way that they do fit a linear pattern. We calculated the logarithm of all the values of both the infant mortality and GDP per capita and plotted the results in the scatterplot below.

c) Describe the relationship between logarithm of infant mortality and logarithm of GDP per capita. Is this relationship linear?
d) The regression line is also included in the scatterplot. Will this regression line be a good prediction of the logarithm of infant mortality given the logarithm of GDP per capita? Why or why not?

End-of-Chapter Exercises

10.CE.1* Answer the following questions on correlation.
   a) Suppose that every student in your class scores 10 points higher on the final exam than on the midterm exam. What would the value of the correlation coefficient between the two exam scores equal? Explain. (Hint: Draw a scatterplot for some hypothetical data.)
   b) Suppose that every student in your class scores 10 points lower on the final exam than on the midterm exam. What would the value of the correlation coefficient between the two exam scores equal? Explain. (Hint: Draw a scatterplot for some hypothetical data.)
   c) Suppose that every student in your class scores twice as many points on the final exam than on the midterm exam. What would the value of the correlation coefficient between the two exam scores equal? Explain. (Hint: Draw a scatterplot for some hypothetical data.)

10.CE.2 Answer the following questions on correlation and regression.
a) When calculating a correlation coefficient between two quantitative variables, does it matter which is considered explanatory and which response?
b) When calculating a least squares line between two quantitative variables, does it matter which is considered explanatory and which response?
c) When calculating a residual value from a least squares line, does it matter whether you calculate the vertical horizontal, or perpendicular distance to the line?

Teacher’s age and distance to school*

10.CE.3 Is there an association between a teacher’s age and how far he/she lives from school? For each of the following descriptions, draw a scatterplot of hypothetical data for a sample of 10 teachers that reveal what’s described. (Be sure to label and put a reasonable scale on both axes.)

a) No association between age and distance
b) Strong, positive, linear association between age and distance
c) Moderate, positive, non-linear association between age and distance
d) Moderate evidence that older teachers tend to live closer to school than younger teachers
e) Very little association between age and distance, except for a severe outlier

Major league baseball

10.CE.4 Suppose that you record the following information about 15 Major League Baseball games played next Saturday:

A. Total number of runs scored in the game
B. Whether or not the home team wins the game
C. Time (in minutes) required to play the game
D. Attendance (number of people) at the game
E. Whether the game is played in the afternoon or evening
F. Temperature when the game begins

a) Identify all pairs of variables for which it does make sense to calculate the correlation coefficient between the variables.
b) Identify all pairs of variables for which it does not make sense to calculate the correlation coefficient between the variables.

(Hint: There should be a total of 15 pairs of variables between your answers to parts a) and b).
Test-Taking Time*

**10.CE.5** Problems 10.CE.5 through 10.CE.7 refer to data on test-taking times. The datafile TestTimes contains data on the time (in minutes) taken by 33 students to complete a multiple choice exam, and the students’ scores (as a percentage out of 100) on the exam. Investigate whether the data provide evidence of an association between the time to take the test and the score achieved on the test.

a) Produce a scatterplot, and describe the association between the variables.
b) Report the value of the correlation coefficient between these variables.
c) Conduct a simulation-based analysis of whether the correlation coefficient differs significantly from zero. Report the p-value, along with a screen capture of the simulated null distribution of correlation values.
d) Conduct a $t$-test based on the correlation coefficient. Report the value of the test statistic and p-value.
e) Summarize your conclusion.

**10.CE.6** Reconsider the previous exercise and the data in TestTimes.

a) Report the equation of the least squares line for reporting test score from time to take the test.
b) Interpret the value of the slope coefficient.
c) Interpret the value of the intercept coefficient. Does this make sense in this context?
d) Report and interpret the value of $r^2$.

**10.CE.7** Reconsider the previous two exercises and the data in TestTimes.

a) Conduct a $t$-test of whether the slope coefficient differs significantly from zero. Report the value of the test statistic and p-value.
b) Produce a 95% confidence interval for the population slope coefficient.
c) Interpret this confidence interval.
d) Are the test result and confidence interval consistent with each other? Explain.
Planets

10.CE.8 Problems 10.CE.8 and 10.CE.9 refer to data gathered on the planets. The following data on the eight planets in our solar system were obtained from Wikipedia:

<table>
<thead>
<tr>
<th></th>
<th>Mercury</th>
<th>Venus</th>
<th>Earth</th>
<th>Mars</th>
<th>Jupiter</th>
<th>Saturn</th>
<th>Uranus</th>
<th>Neptune</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean distance from sun (in million km)</td>
<td>57.9</td>
<td>108.2</td>
<td>149.6</td>
<td>227.9</td>
<td>778.4</td>
<td>1426.7</td>
<td>2871.0</td>
<td>4498.3</td>
</tr>
<tr>
<td>Orbital period (in Earth years)</td>
<td>0.24</td>
<td>0.62</td>
<td>1.00</td>
<td>1.88</td>
<td>11.86</td>
<td>29.45</td>
<td>84.02</td>
<td>164.79</td>
</tr>
</tbody>
</table>

a) Describe the association between these variables as described in the graph on the scatterplot below. Comment on direction, strength, and form of the association.

b) The correlation coefficient between these variables equals 0.988. Would you conclude that the relationship between these variables is linear? Explain.

10.CE.9 Reconsider the previous exercise. The following graph displays the same data, with the least squares line superimposed:

Would you feel comfortable using this line to make predictions? Explain.
Day hikes*

10.CE.10 Problems 10.CE.10 and 10.CE.11 reconsider the data on day hikes in San Luis Obispo county from Exercises 10.3.21 – 10.3.23. The data on hiking time (in minutes) and distance (in miles) can be found in the datafile HikeDistances.

a) Conduct a $t$-test to determine whether the data provide strong evidence of a positive association between hiking time and distance. Report the hypotheses, test statistic, and p-value. Summarize your conclusion.

b) Determine a 95% confidence interval for the population slope coefficient. Interpret this interval, including an interpretation of what slope means in this context.

10.CE.11 Repeat the previous exercise, replacing distance with elevation gain as the explanatory variable. The data on hiking time (in minutes) and elevation gain (in feet) can be found in the datafile HikeElevations.

Walking straight

10.CE.12 While a high school student in Texas, Andrea Axtell conducted a project in which she investigated how well blindfolded students can walk in a straight line. She recruited 30 subjects by randomly selecting students at her school. She put the subjects on the center hash mark of a football field’s goal line, blindfolded them, and asked them to walk in a straight line toward the opposite goal line. She then recorded the yard line at which the subject crossed the sideline, so larger values indicate that the subject walked farther before veering off course. The data, including the heights of the 30 subjects, appear in the file WalkingStraight.

Analyze the data to investigate whether taller people tend to walk farther before veering off course. Write a paragraph summarizing your findings. Include an appropriate graph and numerical summaries. Also conduct inference (p-values, confidence intervals) using both simulation- and theory-based methods. Summarize your conclusions.
Late to class*

10.CE.13 Does having a professor show up late for class, influence students to be late for class? Or when professors are early, do students tend to show up early? In the spring of 2008, some statistics students randomly chose 31 different classes and recorded how many minutes a professor was early or late for class and the average number of minutes the students in the class were early or late. They wanted to see if there was a positive linear relationship between the professor’s and the students’ arrival times. We input these data into statistical software and found the correlation was 0.227. We then shuffled the data 1000 times and found the corresponding 1000 correlations. The simulated correlations are shown in the following graph.

- a) What is the null hypothesis in this study?
- b) What is the alternative hypothesis in this study?
- c) Which variable would it seem is the explanatory variable and which is the response?
- d) Based on the histogram, what is the approximate p-value for this study?
e) What would be your complete conclusion for the one-sided study?

f) If this were a two-sided test, how would the alternative hypothesis change and what would be the new approximate p-value?

g) Even if you conclude that the association between these variables is statistically significant, can you legitimately conclude that the instructor’s being late caused students to be late? Explain.

**Least squares**

10.CE.14 Consider a miniature data set:

\[
\begin{array}{c|c|c|c}
x & 0 & 1 & 2 \\
y & 1 & 0 & 2 \\
\end{array}
\]

Four copies of the scatterplot are shown below, each with a fitted line. Find the residuals for each line, and the sum of squares. Match the sum of squared residuals to the plots, choosing from 3/2, 2, 2, and 9/4.
Noise in New York*

10.CE.15 Problems 10.CE.15 through 10.CE.20 refer to the data on decibel levels in New York. On July 20, 2012 *The New York Times* ran an article about the noise level at 33 different locations in New York City. The scatterplots below are all based on that study. Each shows noise level in decibels (dB) versus time of day (24 hour clock).

![Scatterplots showing noise level versus time of day for different categories.](image-url)
10.CE.15 Match each scatterplot with its coefficient and regression slope, choosing from (a) – (f):

(a) Correlations: 0.08  0.15  0.43  0.66  0.73  0.79.
Slopes: 0.38  1.31  2.29  1.83  3.62  1.64

10.CE.16 Match each scatterplot with its verbal description (a) – (e). You may need to use some descriptions more than once, and others may not be needed.

a) No big outliers, no influential points; strong linear relationship.
b) No big outliers, no influential points; no evidence of a linear relationship.
c) No big outliers, but two influential points. Removing these points will decrease the correlation and increase the fitted slope.

10.CE.17 Fill in the blanks based on the patterns in the data.

a) For restaurants, the correlation between average noise level and time becomes _______ (stronger, weaker) if you exclude the two lunch-time measurements.
b) For stores, the noise level _______ (is, is not) related to time of day.
c) For the “other” locations, the suggestion of a possible relationship between noise level and time of day becomes _______ (stronger, weaker) if you exclude the gym measured at 6:30 pm.
d) On average, restaurants are _______ (more, less) noisy than stores.

e) On average, decibel levels measured later in the day tend to be _______ (higher, lower) than those measured earlier in the day.
f) Time of day _______ (is, is not) confounded with type of location.
g) For _______ (restaurants, stores, other locations) the fitted noise level increases by more than 10 decibels during the three hours from 7 pm to 10 pm.
h) If you go by the fitted line for all restaurants, the predicted decibel level at 4 pm is closest to _______ (round to the nearest 10 dB).

10.CE.18 Refer to description in 10.CE.15.

a) Why is it not possible to conclude from the data that decibel levels increase as it gets later in the day?
b) How could the design of the study be improved in order to allow better conclusions about the effect of time of day on noise level?

**10.CE. 19** Refer to plots in 10.CE.15.

a) Refer to plot A, the scatterplot for all locations in the survey. Note that between 12 noon and midnight (24) the fitted decibel level goes from about 75 to about 95. Use this fact to estimate the decibel level at the y-intercept. (For comparison, a measurement taken in the Evergreens Cemetery was 60 dB.)
b) What time of day does the y-intercept correspond to? Explain why the fitted decibel level for time 0 cannot possibly be right.

**10.CE. 20** Three histograms are shown below. Each shows simulated values of r, obtained by breaking the association on one of the scatterplots (A) – (C) in (10.CE.15). Which histogram is for each plot? How can you tell?
Coffee and height

10.CE.21 Suppose a friend of yours is telling you about a study reported in a popular magazine. The friend said that the study reported a correlation of zero between the amount of coffee a person drank as a child and his or her height as an adult. Your friend interprets this correlation to mean that the more coffee a child drinks, the shorter that person will be as an adult. In other words, drinking coffee stunts one's growth. Explain what is wrong with your friend's interpretation and explain what this type of correlation means.

US States*

10.CE.22 This question is based on the fifty US states. Each state is an observational unit, and the variables listed are measured separately for each of the 50 states. For each of the following pairs of variables, would you expect the correlation to be: strong and positive, moderate and positive, weak or very weak, moderate and negative, or strong and negative.

a) 2012 popular vote for Romney and 2012 popular vote for Obama
b) Total 2012 popular vote for president and 2012 number of electoral votes
c) Number of McDonald’s restaurants and number of attorneys
d) High school graduation rate and percentage of people over 25 with a four-year college degree
e) Year of statehood and number of years as a state
f) Year of statehood and longitude (east/west)
g) Longitude and latitude (north/south)
h) Latitude and average annual temperature

Investigation: Association between hand span and candy?

Is hand span a good predictor of how much candy you can grab? Using 45 college students as subjects, researchers set out to explore whether a linear relationship exists between hand span (cm) and the number of tootsie rolls each subject could grab.

Step 1: Ask a research question

1. State the research question.

Step 2: Design a study and collect data
2. What are the observational/experimental units?

3. Is this a randomized experiment or an observational study? Explain how you know.

4. State the two variables measured on each unit.

5. Is there an explanatory/response relationship for these variables? Classify the variables in this study as categorical or quantitative.

6. State the null and alternative hypotheses to be investigated with this study.

Step 3: Explore the data

7. Use the data set HandSpan to calculate means, standard deviations and the correlation coefficient.

8. Create a scatterplot of the data with hand span on the x axis and number of tootsie rolls on the y axis.

9. Does there appear to be an association between hand span and number of tootsie rolls? Describe the direction, form and strength of the association: is it positive, negative, weak, strong, linear?

10. Are there any unusual observations?

11. Find the least squares regression equation that predicts number of tootsie rolls based on hand span.

12. Interpret the slope of the least squares regression equation in terms of hand span and predicted number of tootsie rolls grabbed.

Step 4: Draw inferences
13. Use the appropriate applet to construct a simulated null distribution using the slope of the least squares regression line as the statistic. Mark the observed least squares slope on this graph.

a) Paste a screen shot or draw your null distribution below with the observed slope marked and the approximate p-value shaded in.

b) Is the observed statistic out in the tail of this null distribution, or is it a fairly typical result?

c) What is the p-value from your simulation? Based on this simulation analysis, would you conclude that the data provides strong evidence against the null hypothesis and conclude that there is a significant association between hand span and number of tootsie rolls grabbed? Explain your reasoning.

14. Are the validity conditions met to complete a theory-based $t$-test on these data? Explain. Whether the conditions are met or not, use a theory-based $t$-test to find a p-value. How does this p-value compare to the p-value your found using the simulation-based method?

15. Find a confidence interval for the slope that would describe the association between all hand spans and number of tootsie rolls grabbed. Is zero in this interval? Does this make sense based on your p-value from your test of significance?

Step 5: Formulate conclusions

16. Are you able to conclude that the association between hand span and number of tootsie rolls grabbed is causal? Explain.

17. What generalizations are you willing to make? Explain.

Step 6: Look back and ahead
18. Summarize the findings from the study. What worked well in this study design? What would you change? What are some follow-up research questions you would like to explore based on the findings from this study?

Research Article: Music preferences and delinquency in youth

Read “Early adolescent music preferences and minor deliquency.” by ter Bogt, Keijzers, and Meeus et al. (2013), 131(2), 1-10 in the journal Pediatrics, and then answer the following questions.

Step 1. Ask a research question
1. In no more than 1-2 sentences state a broad research question (or related questions) that the researchers were trying to investigate in this study.
2. Identify at least two (cited) points of evidence the researchers cite as to why the research question(s) are important. Include the citation in your response.

Step 2. Design a study and collect data
3. How many different adolescents were in the study?
4. What age were the adolescents when the study started?

Step 3. Explore the data
5. Based on Table 1, at age 12, on average, what were the two most liked genres of music? What were their average values?
6. What were the two genres that were least liked at age 12? What were their average values?
7. Answer question #5 for the sample at age 16.
8. Answer question #6 for the sample at age 16.

Step 4. Draw inferences
Table 2 shows the results of fitting regression lines or quadratic regression lines (curves) to the musical preference scores over the four measurements (age 12, 14, 15, 16) in the study.
9. The chart pop musical preference regression line is given as Preference Score = 4.12 - 0.12*(Age-12), where “Age-12” is the age of the participant minus 12 (so, when the
participant is age 12 the value of the variable is 0. **Ignore the quadratic slope coefficient for this question.**

- a. Interpret the intercept.
- b. Interpret the slope
- c. Notice that the slope is significant (p < 0.05, as indicated by an asterisk). State the null and alternative hypothesis, indicate the strength of evidence provided by the p-value and provide an appropriate (in context) interpretation.

10. Answer the same questions as in #9 for the hip-hop genre using the regression line
   
   \[
   \text{Preference Score} = 3.27 + 0.10\times(\text{Age-12})
   \]

Table 3 shows the correlations between numerous variables in the study. For each of the following pairs of variables State the null and alternative hypothesis, indicate the strength of evidence provided by the p-value (p > 0.05 (no asterisk), p < 0.05 (*) or p < 0.01 (**) and provide an appropriate (in context) interpretation.

11. Chart pop preference and R&B preference at age 12 (correlation = 0.44**)
12. Classic preference and Hip-hop preference at age 12 (correlation = -0.22**)
13. Metal preference at age 12 and Delinquency at age 16 (correlation = 0.29**)
14. Chart pop preference at age 12 and Delinquency at age 16 (correlation = -0.07)

**Step 5. Formulate conclusions**

15. To which adolescents are you willing to generalize your conclusions from this study? Why?
16. Does the significant result in question #13 mean that liking the Metal genre at age 12 causes adolescents to be more delinquent at age 16?

**Step 6. Look back and ahead**

17. Identify at least two things that, if you were running the study and had a reasonable amount of financial resources, you would do differently. You must justify your answer.