In this chapter we introduce methods for determining whether a correlation, or association, between two variables exists and whether the correlation is linear. For linear correlations, we can identify an equation that best fits the data and we can use that equation to predict the value of one variable given the value of the other variable. In this chapter, we also present methods for analyzing differences between predicted values and actual values.

In addition, we consider methods for identifying linear equations for correlations among three or more variables. We conclude the chapter with some basic methods for developing a mathematical model that can be used to describe nonlinear correlations between two variables.

Key Concept
In part 1 of this section introduces the **linear correlation coefficient** \( r \), which is a numerical measure of the strength of the relationship between two variables representing quantitative data.

Using **paired sample data** (sometimes called **bivariate data**), we find the value of \( r \) (usually using technology), then we use that value to conclude that there is (or is not) a linear correlation between the two variables.

Key Concept
In this section we consider only **linear relationships**, which means that when graphed, the points approximate a **straight-line pattern**.

In Part 2, we discuss methods of **hypothesis testing for correlation**.
Part 1: Basic Concepts of Correlation

Definitions

A correlation exists between two variables when the values of one are somehow associated with the values of the other in some way.

The linear correlation coefficient $r$ measures the strength of the linear relationship between the paired quantitative $x$- and $y$-values in a sample.

Exploring the Data

We can often see a relationship between two variables by constructing a scatterplot.

Figure 10-2 following shows scatterplots with different characteristics.

Scatterplots of Paired Data

(a) Positive correlation: $r = 0.851$

(b) Negative correlation: $r = -0.965$

Figure 10-2

(c) No correlation: $r = 0$

(d) Nonlinear relationship: $r = 0.807$

Figure 10-2
Requirements

1. The sample of paired \((x, y)\) data is a **simple random** sample of quantitative data.

2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.

3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating \(r\) with and without the outliers included.

Notation for the Linear Correlation Coefficient

\[ n \] = number of pairs of sample data.

\[ \Sigma \] denotes the addition of the items indicated.

\[ \Sigma x \] denotes the sum of all \(x\)-values.

\[ \Sigma x^2 \] indicates that each \(x\)-value should be squared and then those squares added.

\[ (\Sigma x)^2 \] indicates that the \(x\)-values should be added and then the total squared.

\[ \Sigma xy \] indicates that each \(x\)-value should be first multiplied by its corresponding \(y\)-value. After obtaining all such products, find their sum.

\[ r \] = linear correlation coefficient for sample data.

\[ \rho \] = linear correlation coefficient for population data.

Formula

The linear correlation coefficient \(r\) measures the strength of a linear relationship between the paired values in a sample.

**Formula 10-1**

\[
r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2} \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}
\]

Computer software or calculators can compute \(r\).
Interpreting $r$

Using Table A-6: If the absolute value of the computed value of $r$, denoted $|r|$, exceeds the value in Table A-6, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

If $|r| > \text{critical value}$, then there is a linear correlation.

Using Software: If the computed P-value is less than or equal to the significance level, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

If P-value $\leq \alpha$, then there is a linear correlation.

Caution

Know that the methods of this section apply to a \textit{linear correlation}. If you conclude that there does not appear to be linear correlation, know that it is possible that there might be some other association that is not linear.

Properties of the Linear Correlation Coefficient $r$

1. $-1 \leq r \leq 1$
2. If all values of either variable are converted to a different scale, the value of $r$ does not change.
3. The value of $r$ is not affected by the choice of $x$ and $y$. Interchange all $x$- and $y$-values and the value of $r$ will not change.
4. $r$ measures strength of a linear relationship.
5. $r$ is very sensitive to outliers, they can dramatically affect its value.
Example:
The paired pizza/subway fare costs from Table 10-1 are shown here in Table 10-2. Use computer software with these paired sample values to find the value of the linear correlation coefficient \( r \) for the paired sample data.

Requirements are satisfied: simple random sample of quantitative data; Minitab scatterplot approximates a straight line; scatterplot shows no outliers - see next slide

Example:
Using software or a calculator, \( r \) is automatically calculated:

<table>
<thead>
<tr>
<th>MINITAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlations: Pizza, Subway</td>
</tr>
<tr>
<td>Pearson correlation of Pizza and Subway = 0.988</td>
</tr>
<tr>
<td>P-Value = 0.000</td>
</tr>
</tbody>
</table>

Interpreting the Linear Correlation Coefficient \( r \)

We can base our interpretation and conclusion about correlation on a \( P \)-value obtained from computer software or a critical value from Table A-6.

Using Computer Software to Interpret \( r \):
If the computed \( P \)-value is less than or equal to the significance level, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Using Table A-6 to Interpret \( r \):
If \(|r|\) exceeds the value in Table A-6, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.
Interpreting the Linear Correlation Coefficient $r$

| $r$ | > critical value; there is a linear correlation

Example:
Requirements are satisfied: simple random sample of quantitative data; scatterplot approximates a straight line; no outliers

Using Software to Interpret $r$:
The $P$-value obtained from software is 0.364. Because the $P$-value is not less than or equal to 0.05, we conclude that there is not sufficient evidence to support a claim of a linear correlation between weights of discarded paper and glass.

Example:
Using Table A-6 to Interpret $r$:
If we refer to Table A-6 with $n = 62$ pairs of sample data, we obtain the critical value of 0.254 (approximately) for $\alpha = 0.05$. Because $|0.117|$ does not exceed the value of 0.254 from Table A-6, we conclude that there is not sufficient evidence to support a claim of a linear correlation between weights of discarded paper and glass.

Example:
Using a 0.05 significance level, interpret the value of $r = 0.117$ found using the 62 pairs of weights of discarded paper and glass listed in Data Set 22 in Appendix B. When the paired data are used with computer software, the $P$-value is found to be 0.364. Is there sufficient evidence to support a claim of a linear correlation between the weights of discarded paper and glass?
Interpreting $r$: Explained Variation

The value of $r^2$ is the proportion of the variation in $y$ that is explained by the linear relationship between $x$ and $y$.

Using the pizza subway fare costs in Table 10-2, we have found that the linear correlation coefficient is $r = 0.988$. What proportion of the variation in the subway fare can be explained by the variation in the costs of a slice of pizza?

With $r = 0.988$, we get $r^2 = 0.976$.

We conclude that 0.976 (or about 98%) of the variation in the cost of a subway fare can be explained by the linear relationship between the costs of pizza and subway fares. This implies that about 2% of the variation in costs of subway fares cannot be explained by the costs of pizza.

Example:

Common Errors Involving Correlation

1. **Causation**: It is wrong to conclude that correlation implies causality.

2. **Averages**: Averages suppress individual variation and may inflate the correlation coefficient.

3. **Linearity**: There may be some relationship between $x$ and $y$ even when there is no linear correlation.

Caution Know that correlation does not imply causality.

Part 2: Formal Hypothesis Test

We wish to determine whether there is a significant linear correlation between two variables.
Hypothesis Test for Correlation - Notation

\[ n = \text{number of pairs of sample data} \]
\[ r = \text{linear correlation coefficient for a sample of paired data} \]
\[ \rho = \text{linear correlation coefficient for a population of paired data} \]

Hypothesis Test for Correlation - Hypotheses

\[ H_0: \rho = 0 \quad (\text{There is no linear correlation.}) \]
\[ H_1: \rho \neq 0 \quad (\text{There is a linear correlation.}) \]

Test Statistic: \( r \)
Critical Values: Refer to Table A-6

Hypothesis Test for Correlation - Requirements

1. The sample of paired \((x, y)\) data is a simple random sample of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating \( r \) with and without the outliers included.

Hypothesis Test for Correlation - Conclusion

If \(| r | > \text{critical value from Table A-6}\), reject \( H_0 \) and conclude that there is sufficient evidence to support the claim of a linear correlation.

If \(| r | \leq \text{critical value from Table A-6}\), fail to reject \( H_0 \) and conclude that there is not sufficient evidence to support the claim of a linear correlation.
Example:

Use the paired pizza subway fare data in Table 10-2 to test the claim that there is a linear correlation between the costs of a slice of pizza and the subway fares. Use a 0.05 significance level.

Requirements are satisfied as in the earlier example.

\[ H_0: \rho = 0 \quad \text{(There is no linear correlation.)} \]
\[ H_1: \rho \neq 0 \quad \text{(There is a linear correlation.)} \]

Example:

The test statistic is \( r = 0.988 \) (from an earlier Example). The critical value of \( r = 0.811 \) is found in Table A-6 with \( n = 6 \) and \( \alpha = 0.05 \). Because \( |0.988| > 0.811 \), we reject \( H_0: \rho = 0 \). (Rejecting “no linear correlation” indicates that there is a linear correlation.)

We conclude that there is sufficient evidence to support the claim of a linear correlation between costs of a slice of pizza and subway fares.

Hypothesis Test for Correlation - from a \( t \) Test or \( P \)-Value

\[ H_0: \rho = 0 \quad \text{(There is no linear correlation.)} \]
\[ H_1: \rho \neq 0 \quad \text{(There is a linear correlation.)} \]

Test Statistic: \( t \)

\[
 t = \frac{r}{\sqrt{1 - r^2}} \quad \sqrt{n - 2}
\]

Hypothesis Test for Correlation - Conclusion

\( P \)-value: Use computer software or use Table A-3 with \( n - 2 \) degrees of freedom to find the \( P \)-value corresponding to the test statistic \( t \).

If the \( P \)-value is less than or equal to the significance level, reject \( H_0 \) and conclude that there is sufficient evidence to support the claim of a linear correlation.

If the \( P \)-value is greater than the significance level, fail to reject \( H_0 \) and conclude that there is not sufficient evidence to support the claim of a linear correlation.
Example:

Use the paired pizza subway fare data in Table 10-2 and use the $P$-value method to test the claim that there is a linear correlation between the costs of a slice of pizza and the subway fares. Use a 0.05 significance level.

Requirements are satisfied as in the earlier example.

$\textbf{H}_0: \rho = 0$  \hspace{1cm} (There is no linear correlation.)

$\textbf{H}_1: \rho \neq 0$  \hspace{1cm} (There is a linear correlation.)

Example:

The linear correlation coefficient is $r = 0.988$ (from an earlier Example) and $n = 6$ (six pairs of data), so the test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.988}{\sqrt{\frac{1-0.988^2}{6-2}}} = 12.793$$

With df = 4, Table A-6 yields a $P$-value that is less than 0.01. Computer software generates a test statistic of $t = 12.692$ and $P$-value of 0.00022.

Example:

Using either method, the $P$-value is less than the significance level of 0.05 so we reject $\textbf{H}_0: \rho = 0$.

We conclude that there is sufficient evidence to support the claim of a linear correlation between costs of a slice of pizza and subway fares.

One-Tailed Tests

One-tailed tests can occur with a claim of a positive linear correlation or a claim of a negative linear correlation. In such cases, the hypotheses will be as shown here.

<table>
<thead>
<tr>
<th>Claim of Negative Correlation</th>
<th>Claim of Positive Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Left-tailed test)</td>
<td>(Right-tailed test)</td>
</tr>
<tr>
<td>$\textbf{H}_0: \rho = 0$</td>
<td>$\textbf{H}_0: \rho = 0$</td>
</tr>
<tr>
<td>$\textbf{H}_1: \rho &lt; 0$</td>
<td>$\textbf{H}_0: \rho &gt; 0$</td>
</tr>
</tbody>
</table>

For these one-tailed tests, the $P$-value method can be used as in earlier chapters.
Recap

In this section, we have discussed:

- Correlation.
- The linear correlation coefficient $r$.
- Requirements, notation, and formula for $r$.
- Interpreting $r$.
- Formal hypothesis testing.

535/8. Supermodel Height and Weight

The heights and weights of a sample of 9 supermodels were measured. Using a TI-83/84 Plus calculator, the linear correlation coefficient of the 9 pairs of measurements is found to be 0.360. (The supermodels are Alves, Avermann, Hilton, Dyer, Turlington, Hall, Campbell, Mazza, and Hume.) Is there sufficient evidence to support the claim that there is a linear correlation between the heights and weights of supermodels? Explain.

535/9. a. Construct a scatterplot. (S102A; S102B)

b. Find the value of the linear correlation coefficient $r$, then determine whether there is sufficient evidence to support the claim of a linear correlation between the two variables.

c. Identify the feature of the data that would be missed if part (b) was completed without constructing the scatterplot.

x:   10     8     13     9     11     14     6     4     12     7     5

See TI Calculator tutorials for instructions:
http://cfcc.edu/faculty/cmoore/TI83Modeling.htm

535/6. Heights of Mothers and Daughters

The heights (in inches) of a sample of eight mother-daughter pairs of subjects were measured. Using a TI-83/84 Plus calculator with the paired mother-daughter heights, the linear correlation coefficient is found to be 0.693 (based on data from the National Health Examination Survey). Is there sufficient evidence to support the claim that there is a linear correlation between the heights of mothers and the heights of their daughters? Explain.

See TI Calculator tutorials for instructions:
http://cfcc.edu/faculty/cmoore/TI83LinRegTTest.htm
535/10. a. Construct a scatterplot. (S102A; S102C)

b. Find the value of the linear correlation coefficient $r$, then determine whether there is sufficient evidence to support the claim of a linear correlation between the two variables.

c. Identify the feature of the data that would be missed if part (b) was completed without constructing the scatterplot.

$$
\begin{align*}
\text{x:} & \quad 10 & 8 & 13 & 9 & 11 & 14 & 6 & 4 & 12 & 7 & 5 \\
\text{y:} & \quad 7.46 & 6.77 & 12.74 & 7.11 & 7.81 & 8.84 & 6.08 & 5.39 & 8.15 & 6.42 & 5.73
\end{align*}
$$

Testing for a Linear Correlation. In Exercises 13–28, construct a scatterplot, find the value of the linear correlation coefficient $r$, and find the critical values of $r$ from Table A-6 using $\alpha = 0.05$. Determine whether there is sufficient evidence to support a claim of a linear correlation between the two variables. (Save your work because the same data sets will be used in Section 10-3 exercises.)

536/16. Heights of Presidents and Runners-Up

Theories have been developed about the heights of winning candidates for the U. S. presidency and the heights of candidates who were runners-up. Listed below are heights (in inches) from recent presidential elections. Is there a linear correlation between the heights of candidates who won and the heights of the candidates who were runners-up? *(S102F; S102G)*

**Winner:** 69.5 73 73 74 74.5 74.5 71 71

**Runner-Up:** 72 69.5 70 68 74 74 73 76

Testing for a Linear Correlation. In Exercises 13–28, construct a scatterplot, find the value of the linear correlation coefficient $r$, and find the critical values of $r$ from Table A-6 using $\alpha = 0.05$. Determine whether there is sufficient evidence to support a claim of a linear correlation between the two variables. (Save your work because the same data sets will be used in Section 10-3 exercises.)

537/20. Commuters and Parking Spaces

Listed below are the numbers of commuters and the numbers of parking spaces at different Metro-North railroad stations (based on data from Metro-North). Is there a linear correlation between the numbers of commuters and the numbers of parking spaces? *(S102H; S102I)*

**Commuters:** 3453 1350 1126 3120 2641 277 579 2532

**Parking Spaces:** 1653 676 294 950 1216 179 466 1454

Testing for a Linear Correlation. In Exercises 13–28, construct a scatterplot, find the value of the linear correlation coefficient $r$, and find the critical values of $r$ from Table A-6 using $\alpha = 0.05$. Determine whether there is sufficient evidence to support a claim of a linear correlation between the two variables. (Save your work because the same data sets will be used in Section 10-3 exercises.)

**CPI and Subway Fare**

The paired values of the Consumer Price Index (CPI) and the cost of subway fare from Table 10-1 in the Chapter Problem are listed below. Is there a linear correlation between the CPI and subway fare? *(S102D; S102E)*

<table>
<thead>
<tr>
<th>CPI</th>
<th>Subway Fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>30.2</td>
<td>48.3</td>
</tr>
<tr>
<td>112.3</td>
<td>162.2</td>
</tr>
<tr>
<td>191.9</td>
<td>197.8</td>
</tr>
</tbody>
</table>

Testing for a Linear Correlation. In Exercises 13–28, construct a scatterplot, find the value of the linear correlation coefficient $r$, and find the critical values of $r$ from Table A-6 using $\alpha = 0.05$. Determine whether there is sufficient evidence to support a claim of a linear correlation between the two variables. (Save your work because the same data sets will be used in Section 10-3 exercises.)
Testing for a Linear Correlation. In Exercises 13–28, construct a scatterplot, find the value of the linear correlation coefficient \( r \), and find the critical values of \( r \) from Table A-6 using \( \alpha = 0.05 \). Determine whether there is sufficient evidence to support a claim of a linear correlation between the two variables. (Save your work because the same data sets will be used in Section 10-3 exercises.)

537/22. New Car Mileage Ratings Listed below are combined city–highway fuel economy ratings (in mi/gal) for different cars. The old ratings are based on tests used before 2008 and the new ratings are based on tests that went into effect in 2008. Is there sufficient evidence to conclude that there is a linear correlation between the old ratings and the new ratings? (S102J; S102K)

<table>
<thead>
<tr>
<th>New</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>16</td>
<td>27</td>
<td>17</td>
<td>33</td>
<td>28</td>
<td>24</td>
<td>18</td>
</tr>
</tbody>
</table>

539/24. Costs of Televisions Listed below are prices (in dollars) and quality rating scores of rear-projection televisions (based on data from Consumer Reports). All of the televisions have screen sizes of 55 in. or 56 in. Is there sufficient evidence to conclude that there is a linear correlation between the price and the quality rating score of rear-projection televisions? Does it appear that as the price increases, the quality score also increases? Do the results suggest that as you pay more, you get better quality? (S102L; S102M)

<table>
<thead>
<tr>
<th>Price</th>
<th>2300</th>
<th>1800</th>
<th>2500</th>
<th>2700</th>
<th>2000</th>
<th>1700</th>
<th>1500</th>
<th>2700</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality Score</td>
<td>74</td>
<td>73</td>
<td>70</td>
<td>66</td>
<td>63</td>
<td>62</td>
<td>52</td>
<td>68</td>
</tr>
</tbody>
</table>

539/26. Crickets and Temperature One classic application of correlation involves the association between the temperature and the number of times a cricket chirps in a minute. Listed below are the numbers of chirps in 1 min and the corresponding temperatures in °F (based on data from The Song of Insects by George W. Pierce, Harvard University Press). Is there a linear correlation between the number of chirps in 1 min and the temperature? (S102N; S102O)

<table>
<thead>
<tr>
<th>Chirps in 1 min</th>
<th>882</th>
<th>1188</th>
<th>1104</th>
<th>864</th>
<th>1200</th>
<th>1032</th>
<th>960</th>
<th>900</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature (° F)</td>
<td>69.7</td>
<td>93.3</td>
<td>84.3</td>
<td>76.3</td>
<td>88.6</td>
<td>82.6</td>
<td>71.6</td>
<td>79.6</td>
</tr>
</tbody>
</table>

Testing for a Linear Correlation. In Exercises 13–28, construct a scatterplot, find the value of the linear correlation coefficient \( r \), and find the critical values of \( r \) from Table A-6 using \( \alpha = 0.05 \). Determine whether there is sufficient evidence to support a claim of a linear correlation between the two variables. (Save your work because the same data sets will be used in Section 10-3 exercises.)