Chapter Summary

- A **scatterplot** is a graph that displays the relationship between two numerical variables.
- In a scatterplot, the **explanatory variable** is plotted on the horizontal axis and the **response variable** is plotted on the vertical axis. The response variable is usually the one we are more interested in—either because we want to predict the response for a particular value of the explanatory variable or because we want to use the explanatory variable to explain changes in the response variable.
- Two variables have an **association** if specific values of one variable tend to occur in common with specific values of the other. To describe the association in a scatterplot, it is important to address the direction, form, and strength of the association, along with any unusual values.
- There are three ways to describe the **direction** of an association:
  - Two variables have a **positive association** if large values of one variable are typically paired with large values of the other variable and small values are paired with small values.
  - Two variables have a **negative association** if large values of one variable are typically paired with small values of the other variable.
  - Two variables have **no association** if knowing the values of one variable does not give any useful information about the values of the other variable.
- The **form** of an association is **linear** if the pattern of the points is best described by a straight line. Otherwise, the form is **nonlinear.**
- The **strength** of an association describes the amount of scatter from the overall form of the data. In a strong association, there isn't much scatter and predictions of the response variable will be fairly precise.
- The **correlation (*r*)** is a measure of the strength and direction of a linear association between two numerical variables. Some important characteristics of the correlation include:
  - $-1 \leq r \leq 1$.
  - If the association is negative, then $r < 0$. If the association is positive, then $r > 0$.
  - If there is very little scatter from a linear form, then r is close to 1 or -1. If there is lots of scatter from a linear form, then r is close to 0.
- Even if there is a strong correlation between two numerical variables, it isn't a good idea to conclude that changes in one variable will cause changes in the other variable.
- Because unusual values can have a big influence on the value of the correlation, it is important to make a scatterplot and identify potentially **influential points.**
- The **true correlation** between two variables, like the *ABILITY* of an athlete, exists only in theory. The **observed correlation** between two variables, like the *PERFORMANCE* of an athlete, is based on a limited amount of data, such as one season. Because it is based on a limited amount of data, the observed correlation will vary from the true correlation because of *RANDOM CHANCE*.

- To investigate whether there is convincing evidence that the true correlation between two numerical variables is positive (or negative), we conduct a hypothesis test where the null hypothesis is that the true correlation between the two variables is 0.
- To simulate the distribution of the correlation, shuffle the values of one of the variables and randomly pair them with the values of the other variable. Then calculate the correlation to see what values of the correlation could arise due to *RANDOM CHANCE* alone.
- A **time plot** of a numerical variable plots each *PERFORMANCE* against the time at which it was measured in order to observe possible trends over time and departures from these trends. The time periods are placed on the horizontal axis of the graph and the variable being investigated is placed on the vertical axis.
- A **moving average** is the average of an athlete's *PERFORMANCES* in a specified time period and the time periods immediately before and after the specified time period. Moving averages are used to smooth time plots and make it easier to see trends.