Chapter 5

Using Correlation to Detect Statistical Relationship

5.1 Introduction

One of the primary goals of analytic methods is to understand the relationship between variables. For instance, we might be interested in the relationship between player performance and team performance or in the relationship between performance in a given year and performance in the following year. In this chapter, several different approaches for measuring the strength of the relationship between variables are presented; these measures reduce the properties of such a relationship to a single number that is useful as a single summary of the relationship between variables.

5.2 Linear Relationships: The Correlation Coefficient

Suppose that for each subject, such as a player or team, we measure two variables X and Y and suppose that we are interested in the relationship between these variables. For instance, if the subjects are MLB players, for each player we might record his number of hits (X) and his number of runs recorded (Y) in a given season. The simplest way to gain some insight into the relationship between X and Y is to construct a scatter-plot, in which the pair (X, Y) is plotted for each subject. Figure 5.1 contains a plot of runs scored versus hits for 2011 MLB players with a qualifying number of plate appearances (502 or more); there are 145 such players, so the sample size is $n = 145$.

Like the plot in Figure 5.1, scatter-plots often show a general linear relationship between the variables. In some cases, the pattern is vague, with considerable variation around the underlying linear trend; in other cases, the data values might come close to falling on a specific line. That is, in some cases, the relationship is strong so that the value of X almost completely determines the value of Y; in other cases, the relationship is weak, with the value of X giving, at best, a general indication regarding the value of Y. For instance, the relationship between runs and hits, as given in Figure 5.1, is a fairly strong linear relationship, and the relationship between walks and hits for qualifying 2011 MLB players shown in Figure 5.2, shows only a weak linear relationship between the variables.

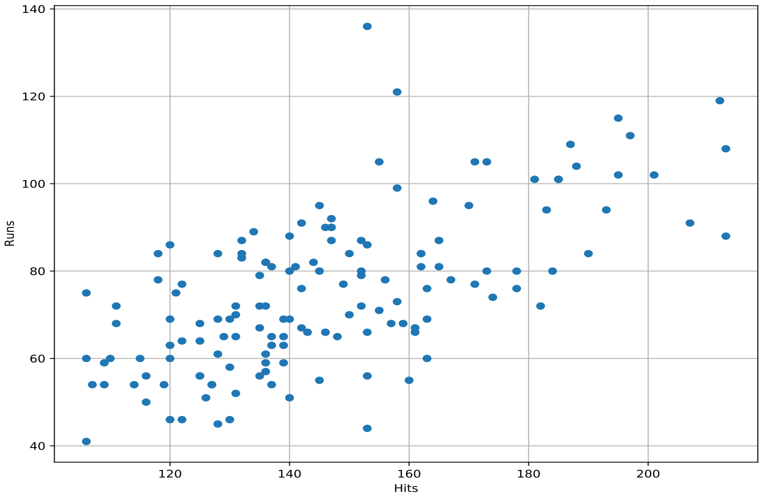

Figure 5.1: Runs versus Hits for 2011 MLB players.

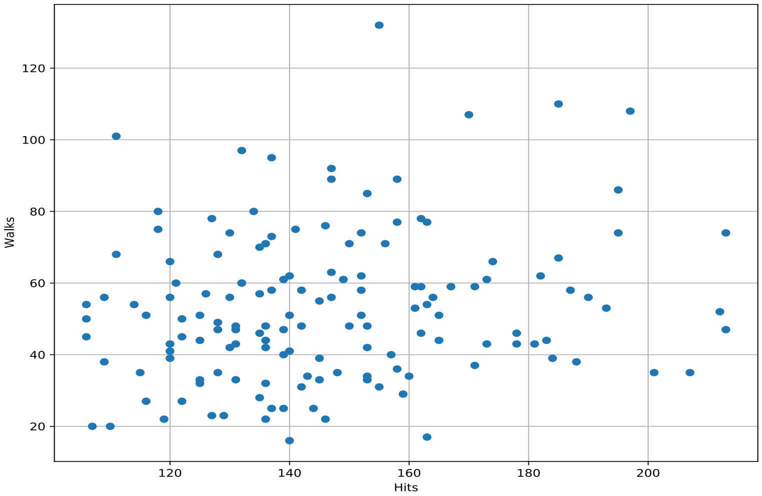

Figure 5.2: Walks versus Hits for 2011 MLB players.

The correlation coefficient, denoted by r, is a single-number measure of the extent to which data cluster around a line. For instance, for runs and hits, as shown in Figure 5.1, $r = 0.64$; for walks and hits, as shown in Figure 5.2, $r = 0.17$. Figures 5.3 to 5.6 give several examples of scatter-plots. Figure 5.3 contains a plot of offensive rebounds per game versus defensive rebounds per game for NBA players with at least 56 games played in 2011-2012 season; here $r = 0.69$ ($n = 182$). Figure 5.4 contains a plot of touch-down passes per passing attempt versus sacks per passing attempt for 2009 NFL season quarterbacks with at least 160 passing attempts; here $r = 0.41$ ($n = 30$). Figure 5.5 contains a plot of 2011 wins versus 2010 wins for MLB team; here

$r = 0.41$ ($n = 30$). Note that the larger the value of $|r|$, the more closely the variables follow a linear relationship, with the sign of r indicating the direction of the relationship.

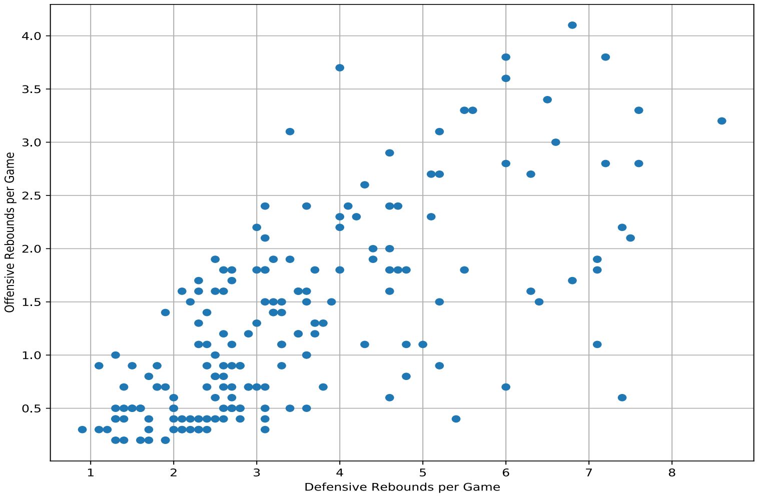

Figure 5.3: Offensive versus defensive rebounds for 2011-2012 NBA players.

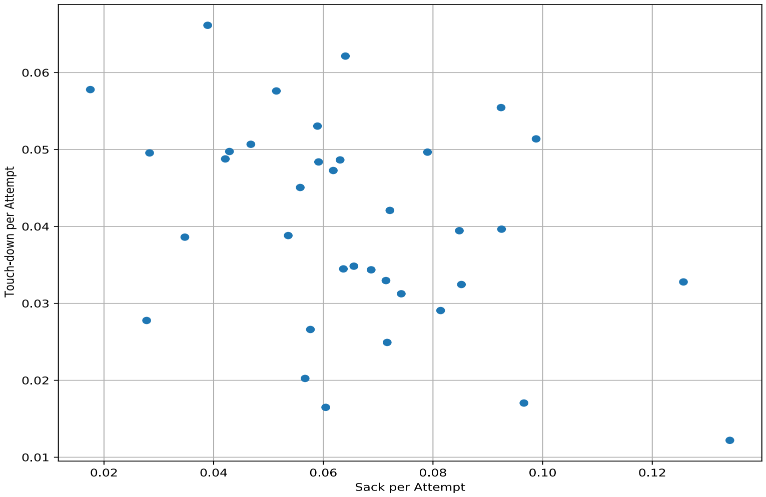

Figure 5.4: Touch-downs versus sacks for 2009 NFL quarterbacks.

The key to effectively using the correlation coefficient is to understand its properties. Let r denote the correlation coefficient. Then,

Correlation Coefficient

- $-1 \leq r \leq 1$
- $r > 0$ indicates an increasing linear relationship in which larger values of X are associated with larger values of Y.
- $r < 0$ indicates a decreasing linear relationship in which larger values of X are associated with smaller values of Y.
- $r = 0$ indicates that there is no linear relationship between the variables.
- $r = 1$ or $r = -1$ indicates a perfect linear relationship between the variables.
- r is affected by switching the roles of the variables.
- Adding a constant to a variable does not change the value of r.
- r is not affected by changing the units of the variables.

Important

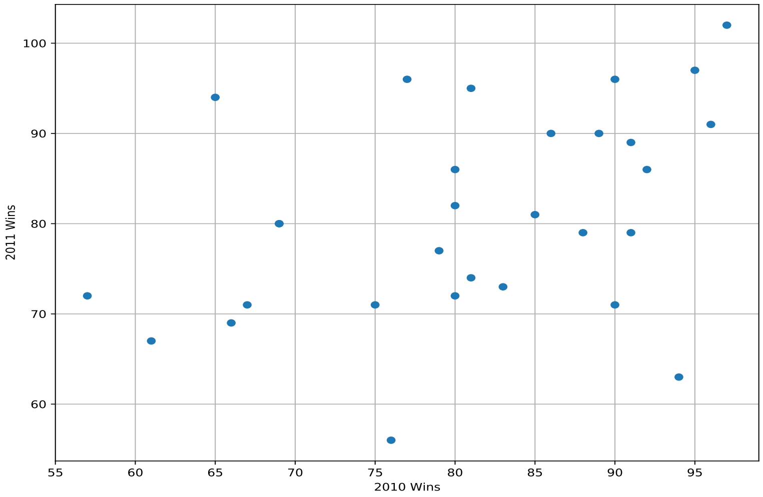

Figure 5.5: 2011 wins versus 2010 wins for MLB teams.

Therefore, values of r near 0 indicate a weak linear relationship, if any at all, with larger values of $|r|$ indicating stronger linear relationships. In particular, if the random variables under consideration are independent, we expect r to be close to 0; it is unlikely to be exactly 0 because of the random nature of the data. Assessments of the magnitude of $|r|$ generally depend on the nature of the variables being considered. In well controlled scientific experiments, values of $|r|$ close to 1 might be observed; however, with sports data, values of $|r|$ greater than 0.9 are relatively rare.

Interest often centers on whether two variables have a linear relationship. Because a value of r of exactly 0 is very unlikely with real data, it is natural to ask if a non-zero correlation coefficient is statistically significant different from 0. A simple way to assess statistical significance is to compare

$$|r| \text{ to } \frac{2}{\sqrt{n}}$$

Wether the Correlation is significant or not.

where n denotes the number of subjects under consideration. If $|r| < 2/\sqrt{n}$, we conclude that, even though it is not exactly 0, there is not a statistically significant linear relationship between the variables. That is, the observed linear relationship may be attributed to chance.

When evaluating significance in this way, there are some important issues to keep in mind; although we discuss these in the context of correlation, they apply to statistical significance in general. Suppose that the sample size is very large; then, an observed correlation can be small but still statistically significant. For example, if $n = 5000$, a correlation as small as 0.03 is still statistically significant. This means that, if the true correlation is actually 0, it is unlikely that we would observe a correlation with magnitude as large as 0.03. However, such a small correlation is unlikely to be practically important. On the other hand, if n is small even relatively large correlations may not be statistically significant. For instance, if $n = 20$, even a correlation as large as 0.44 is not statistically significant. In these cases, it may be that there is a true correlation that is large enough to be important, but the sample size is too small to be able to say that with certainty. If possible, collecting more data will allow us to reduce this uncertainty; otherwise, we can use the observed correlation in the analysis but use caution in drawing firm conclusions. The margin of error of a correlation coefficient r, based on a sample of size n, is given by

$0.44 < \frac{2}{\sqrt{20}}$ → not statistically significant

R: Correlation coefficient

$$\text{ME} \rightarrow 2\sqrt{\frac{1-r^2}{n}} \tag{5.1}$$

margin error

→ it's a good start point but it's not everything

For example, for the correlation between the number of hits and number of runs of MLB players, shown at the beginning of this section, $r = 0.64$, with a margin of error of 0.13. One of the appeals of the correlation coefficient is its simplicity; it reduces the possibly complex relationship between two variables to a single number. But, such simplicity has some drawbacks; hence, there are some important issues to keep in mind when basing conclusions on correlations.

One is that the correlation coefficient considers only one type of relationship, a linear one. Therefore, a correlation coefficient near 0 does not imply that the variables are not related, only that there is no evidence of a linear relationship. Although in some fields non-linear relationships arise naturally and are common, they are relatively uncommon when analyzing sports data. Therefore, although a small value of $|r|$ often indicates that the variables are not related, to be safe, one should always supplement the correlation coefficient with a scatter-plot to rule out the possibility of an important non-linear relationship.

Another important fact is that even if X and Y are correlated; that is, they have a statistically significant correlation, that does not imply that X and Y have any type of *cause-and-effect* relationship. In particular, an observed linear between Y and X might be because both X and Y are related to a third variable.

Correlation $\nrightarrow$ cause and effect

Consider NFL teams during the 2010 and 2011 seasons. Let X denote the draft order of the team for the 2011 draft sot that, for Carolina, $X = 1$ (they took Cam Newton), and for the Super Bowl champion Green Bay Packers, $X = 32$. Let Y denote the number of wins in the 2011 season, the season following the 2011 draft. Then, the correlation of X and Y is 0.36, suggesting that having later draft picks corresponds to more wins in the following season. If the relationship is a causal one, then team with early first-round draft picks should trade for picks later in the round.

Of course, such a causal relationship between draft order and wins in the following season is unlikely; a more likely explanation is that both variables are affected by a third variable, the number of wins in 2010. If a team has few wins in 2010, that team has a high draft pick, and we expect that the team will not win many games in 2011 (although they may win more than in 2010). Let Z be the number of wins in 2010. Because both X and Y are related to Z, they are related to each other. This type of *"lurking variable"* is discussed in Section 5.4.

Finally, it is important to keep in mind that the correlation coefficient is only a single number; therefore, often it does not tell the whole story regarding the relationship between the variables, even when the relationship is a linear one. For instance, consider Figure 5.3 on the relationship between offensive and defensive rebounds of NBA players. The correlation is 0.69, which accurately describes a reasonably strong linear relationship between variables. However, another interesting aspect of the plot of the data is that there is much less variability in the relationship between the variables when both values are small than when both variables are large. That is, players with few defensive rebounds almost always have few offensive rebounds; on the other hand, players with many defensive rebounds not only often have many offensive rebounds but also often have relatively few offensive rebounds. This is an important fact regarding the relationship between offensive and defensive rebounds that is not addressed by simply looking at the correlation; that is, the correlation coefficient is no substitute for examining a plot of the data.

5.3 Using Rank Correlation for Certain Types of Non-Linear Relationships

One drawback of the correlation coefficient is that it measures a specific type of relationship, a linear one between the two variables. Although such linear relationships are common, in some cases we are interested in detecting *some* type of relationship between variables, without regard to its specific form. This is often true if the concept in mind does not involve particular variables but rather some more general type of association.

In these cases, rank correlation is often useful. Like the standard correlation coefficient discusses in the previous section, the rank correlation coefficient is based on two variables measured for each of n subjects. However, instead of analyzing the numerical values of the variables, we first convert the variables to ranks by computing each subject's rank for each variable. For instance, if $n = 3$ and the pairs of measurements for the 3 subjects are

$$(12, 6) \quad (4, 9) \quad (8, 8)$$

The corresponding ranks are

12/3 , 4

$$(1,3) \quad (3,1) \quad (2,2)$$

because, for example, 12 is the largest value of the first variable (rank = 1) and 6 is the smallest value of the second variable (rank = 3). If there are ties, then all subjects involved in the tie receive the average of the ranks involved.

Once the data are converted to ranks, we compute the correlation coefficient using the ranks; the result is denoted by r_s and it is often called *Spearman's rank correlation coefficient* or, simply, the rank correlation coefficient. The rank correlation coefficient has many of the properties of the standard correlation coefficient. For instance, rank correlation always lies between -1 and 1. However, the values 1, -1 no longer indicate a perfect *linear* relationship between the variables, but instead indicate a perfect *monotone* relationship. A monotone relationship is one that is strictly increasing or strictly decreasing. For instance, suppose X is a variable that takes only positive values and suppose $Y = X^2$ exactly. Then, $Y_1 > Y_2$ if and only if $X_1 > X_2$, so that the ranks of the Y values will be exactly the same as the ranks of the X values; in this case, r_s will be exactly 1.

Thus, the properties of the rank correlation coefficient are essentially the same as those of the standard correlation coefficient, except that the linear relationship underlying the standard correlation coefficient are replaced by monotone relationships. For instance, $r_s = 1$ or -1 indicates a perfect monotone relationship between the variables, and $r_s = 0$ indicates that there is no monotone relationship between the variables. Also, the statistical significance of a non-zero value of r_s can be assessed by comparing it to $2/\sqrt{n}$.

5.4 Recognizing and Removing the Effect of a Lurking Variable

As noted in Section 5.2, an observed correlation between two variables sometimes can be explained by the fact that those two variables are both related to a third variable (said to be a *lurking variable*). In some of these cases, the explanation based on the "third variable" might give a better understanding of the mechanism generating the data than does the original correlation.

Very High

For example, consider data on 2009 MLB pitchers with at least 40 innings pitches ($n = 393$). The correlation coefficient for "hits allowed" and "walks allowed" is 0.77. It is tempting to try to explain this correlation by the theory that pitchers yielding more walks have poor control so they make bad pitches more often, which leads to hits. However, an alternative explanation is that pitchers with a lot of innings pitched naturally give up more hits and more walks. Therefore, the observed high correlation between walks allowed and hits allowed might simply be a consequence of the fact that, for this dataset, innings pitched has considerable variation, and both walks allowed and hits allowed are closely related to innings pitched.

In this example, there is a simple remedy: Instead of analyzing total walks and hits allowed, we could analyze hits allowed per 9 innings pitched and walks allowed per 9 innings pitched. In fact, the correlation coefficient for those variables is 0.0024; that is, there is essentially no linear relationship between hits allowed per 9 innings and walks allowed per innings. Therefore, the explanation that the original high correlation between hits allowed and walks allowed can be attributed to innings pitched seems reasonable and appears to be a better explanation that the one based on control issues.

It is not always possible to use a simple standardization like the one based on innings pitched to control for a third variable. In this section, we present a more general approach to the "*lurking variable*" effect, as it is commonly referred.

Let X, Y denote the variables of interest and let Z denote a third variable that might be at least partly responsible for the linear relationship. Let r_{xy} denote the (standard) correlation coefficient for X and Y. Let r_{xz} denote the correlation coefficient for X, Z and let r_{yz} denote the correlation coefficient for Y, Z. It may be shown that, if the correlation for X, Y is entirely a result of their linear relationships with Z, then

$$r_{xy} = r_{xz} r_{yz} \tag{5.2}$$

Therefore, $r_{xy} - r_{xz} r_{yz}$ is a measure of the correlation of X, Y beyond what can be explained by Z. The *partial correlation coefficient* of X, Y controlling for Z is defined by

$$r_{xy,z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \tag{5.3}$$

This partial correlation coefficient "controls for Z" in the following sense: suppose that all variable relationships are linear ones, and the relationship between Y, X is the same for all values of Z. Then, $r_{xy,z}$ measures the correlation between X, Y for a sub-population of subjects all having the same values of Z, that is, holding Z constant. Stated another way, $r_{xy,z}$ represents an estimate of the correlation between X and Y that would be observed if we were somehow able to observe a sample of subjects all having the same value of Z.

Consider the application of these ideas to the example of hits allowed and walks allowed. Let X denote hits allowed, Y denote walks allowed, and Z denote innings pitched. We have seen that $r_{xy} = 0.77$; further analysis shows that $r_{xz} = 0.96$ and $r_{yz} = 0.80$. Note that, to two significant figures, $r_{xy} = r_{xz} r_{yz}$ holds exactly; using exact values shows that $r_{xy,z} = 0.00027$. That is, controlling for innings pitched, there is no linear relationship between hits allowed and walks allowed, essentially the same conclusion we reached using walks and hits per 9 innings.

The assumption underlying partial correlation are important; hence, it is worth considering them in the context of the example. The first assumption is that the relationships between the variables are all linear ones; this can be addressed by looking at the usual scatter-plot of the variables. The second assumption is that the relationship between X, Y is the same for all values of Z. That is, the relationship between hits allowed and walks allowed is the same for all values of innings pitched. For instance, the relationship is the same for starting pitchers, with 200 or more innings pitched, as it is for closers, with less than 100 innings pitched. Although this assumption is unlikely to be exactly true, it does not seem unreasonable; hence, the conclusion that, controlling for innings pitched, there is no relationship between hits allowed and walks allowed is warranted.

Partial correlation coefficients have the same general properties as standard correlation coefficients, except that they measure linear association controlling for the third variable. For example, $r_{xy,z} = 0$ indicates no linear relationship between X, Y controlling for Z.

It is important to note that the analyst is free to choose the controlling variable that seems appropriate in the analysis; different choices lead to different partial correlations. All such choices are valid, provided that the assumptions discussed previously seem reasonable. These different choices lead to partial correlation coefficients that describe different aspects of the relationship between the variables.

5.5 Measures of Association for Categorical Variables

The measures of the strength of the relationship between two variables that have considered so far in this chapter apply to quantitative variables. However, in some cases, the variables of interest are categorical; in this section, we consider methods of measuring the degree of association between such variables. In particular, we focus on the case in which each categorical variable takes two possible values. Similar methods are available for more general categorical variables; however, a number of additional issues arise, making that case beyond the scope of this course.

Consider two categorical variables X and Y and suppose that each variable takes two possible values. For instance, if X denotes the handedness of an MLB pitcher, then X can take the value "L" for left-handed or the value "R" for right-handed. For simplicity, we denote the possible values of each variable as 0 and 1, where the meaning of these values will depend on the variable. For instance, in the example mentioned, we could denote a left-handed pitcher by $X = 0$ to denote a right-handed pitcher by $X = 1$. Alternatively, we could use $X = 0$ to denote a right-handed pitcher and $X = 1$ to denote a left-handed pitcher. Clearly, the conclusions of the analysis should not depend on how we assign the possible values of X to the values 0 and 1.

Let us look at the following example: For each NFL team's starting quarterback in the 2012 season, consider two variables, one that denotes whether the quarterback was a top 10 pick in the draft and a second that denotes whether the quarterback's team made the playoffs. Here, "starting quarterback" is defined to be the quarterback with the most passing attempts for that team during the regular season; in the case of the San Francisco 49ers, in which both Alex Smith and Colin Kaepernick had 218 attempts, Kaepernick was used as the starter. Therefore, there are 32 data points, corresponding to the 32 starting quarterbacks in the NFL. Table 5.1 contains the data for this example.

To measure the association between these variables, we might consider recoding the data so that the correlation coefficient can be applied. Let X denote the variable representing whether the quarterback was a top 10 pick in the draft and let Y denote the variable representing whether the quarterback's team made the playoffs. To calculate the correlation, we can assign values "Yes" and "No" for each variable. For instance, let $X = 1$ if the quarterback was a top 10 pick and let $X = 0$ if he was not a top 10 pick; let $Y = 1$ if the quarterback's team made the playoffs and let

$Y = 0$ if it did not. The correlation between X and Y is -0.119, indicating that quarterbacks who were top 10 picks are slightly less likely to make the playoffs, or stated another way teams that made the playoffs are slightly less likely to have a quarterback who was a top 10 pick.

Table 5.1: Draft status and season result for NFL starting quarterback in 2012

		Team Made the Playoffs		
		No	Yes	Total
Top 10 pick	No	6	5	11
	Yes	14	7	21
	Total	20	12	

Consider a generic table of data, as given in Table 5.2.

Table 5.2: A generic table representing two categorical variables

		Variable 2		
		No	Yes	Total
Variable 1	No	a	b	$a+b$
	Yes	c	d	$c+d$
	Total	$a+c$	$b+d$	

Then, coding Yes as 1 and No as 0, an expression for the correlation coefficient is given by

Correlation coefficient Formula

$$r = \frac{ad - bc}{\boxed{\sqrt{(a+b)(c+d)(a+c)(b+d)}}} \tag{5.4}$$

It is worth noting that, because the correlation coefficient is unaffected by linear transformations of the variables, the correlation is unaffected by the actual values used to represent the different values of the variables. For instance, if Yes for a top 10 pick is given the value 5 and No is given the value 2 and if Yes for making the playoffs is given the value 10 and No is given the value 6, the correlation is still -0.119. However, if the coding changes the order of the categories, the sign of the correlation may change. For instance, if Yes and No for top 10 pick are given the values 0 and 1, respectively, while Yes and No for making the playoffs are given the values 1 and 0, respectively, then the correlation is 0.119.

The correlation coefficient, as applied to categorical data in this way, retains many of the properties that it has when applied to quantitative data. For instance, recall that a correlation of 0 for quantitative data corresponds to the case of no linear relationship between variables. In the case of binary categorical variables, an even stronger result holds: A correlation of 0 implies that the variables are independent.

However, there is an important difference between the correlation coefficient for categorical data and the correlation coefficient for continuous data. While, in general, the correlation takes values in the range -1 to 1, when applied to categorical data, the range of the correlation might be restricted by the distributions fo the two variables under consideration.

To illustrate this, consider the following example: Suppose we are interested in measuring the association between a starting pitcher's league and the probability that he pitches a complete game. Using data from 2012 MLB season, there are 4860 starts, 2952 by National League pitchers and 2268 by American League pitchers; there were 128 complete games. Therefore, without knowing how many National League or American League pitchers pitched complete games, the data are of the form given in Table 5.3.

Table 5.3: Hypothetical data on 2012 MLB starters

		Complete Game		
		No	Yes	Total
League	National			2592
	American			2268
	Total	4732	128	

The correlation between league and complete games will take its maximum value of all the complete games are in one league. For instance, if all 128 complete games are thrown by American League pitchers, the data table will be the one given in Table 5.4. The correlation for this table is 0.176; if all the complete games are thrown by National League pitchers, the correlation is -0.154. Therefore, when studying the relationship between complete games and league, the value of the correlation coefficient must fall in the range -0.154 to 0.176. The actual data for this analysis is given in Table 5.5, and the correlation between league and complete games is 0.024, indicating, at most, a weak relationship between league and complete games.

Table 5.4: Hypothetical data on 2012 MLB starters

		Complete Game		
		No	Yes	Total
League	National	2592	0	2592
	American	2140	128	2268
	Total	4732	128	

One reason for this behavior is that the primary interpretation of the correlation as a measure of "how closely the data cluster around a line" does not really apply in the categorical case. Hence, it is often preferable to use a measure of association designed for categorical variables. To do this, it is helpful to think about what it means for two categorical variables to be associated. Consider variables X and Y, each of which takes the values 0 and 1. These variables are associated if $Y = 1$ occurs relatively more (or less) frequently when $X = 1$ than it does when $X = 0$.

Table 5.5: Actual data on 2012 MLB starters

		Complete Game		
		No	Yes	Total
League	National	2533	59	2592
	American	2199	69	2268
	Total	4732	128	

Note that this is a statement about conditional probabilities or, equivalently, about conditional odds ratios. Let $P(Y = 1|X = 0)$ denote the conditional probability that $Y = 1$ given that $X = 0$ and let $P(Y = 1|X = 1)$ denote the conditional probability that $Y = 1$ given that $X = 1$. The odds of $Y = 1$ versus $Y = 0$ when $X = 0$ are given by the ratio

odds_0

$$\frac{P(Y = 1|X = 0)}{P(Y = 0|X = 0)}$$

similarly, the odds of $Y = 1$ versus $Y = 0$ when $X = 1$ are given by the ratio

odds_1

$$\frac{P(Y = 1|X = 1)}{P(Y = 0|X = 1)}$$

Then, X and Y are associated if

$$\frac{P(Y = 1|X = 0)}{P(Y = 0|X = 0)} \neq \frac{P(Y = 1|X = 1)}{P(Y = 0|X = 1)} \tag{5.5}$$

odds_0 ≠ odds_1 → it's associated

Of course, in practice, we have data, not probabilities. Note that the empirical version of $P(Y = 1|X = 0)$ is the proportion of times $Y = 1$ occurs, restricting attention to only those cases in which $X = 0$ occurs. For data in the form of Table 5.6, this value is $c/(a + c)$. Similarly, the empirical version of $P(Y = 0|X = 0)$ is $a/(a + c)$. It follows that the empirical version of the odds ratio is

$$\frac{P(Y = 1|X = 0)}{P(Y = 0|X = 0)} = \frac{c/(a + c)}{a/(a + c)} = \frac{c}{a}$$

Similarly, the empirical version of the odds ratio

$$\frac{P(Y=1|X=1)}{P(Y=0|X=1)} = \frac{d/(b+d)}{b/(b+d)} = \frac{d}{b}$$

Table 5.6: Generic data for two categorical variables

		X		
		0	1	Total
Y	0	a	b	$a+b$
	1	c	d	$c+d$
	Total	$a+c$	$b+d$	

$P(y=1|x=0)$

$$\frac{P(y=1 \text{ and } x=0)}{P(x=0)}$$

$$\frac{c}{a+c}$$

$$\frac{\frac{c}{\cancel{a+b+c+d}}}{\frac{a+c}{\cancel{a+b+c+d}}}$$

To compare the relative likelihood of $Y = 1$ when $X = 1$ and when $X = 0$, we can look at one odds ratio divided by the other, that is, the ratio of the odds ratio,

$$\frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=0)/P(Y=0|X=0)}$$

This quantity is greater than 1 if $Y = 1$ is relatively more likely when $X = 1$ than when $X = 0$; conversely, it is less than 1 if $Y = 1$ is relatively less likely when $X = 1$ than when $X = 0$. The empirical version of the ratio of odds ratio is

$$\frac{ad}{bc}$$

this is known as the *cross-product ratio* of the data shown in Table 5.6. Note that if we switch the roles of X and Y in the analysis, so that we are comparing the likelihood of $X = 1$ when $Y = 1$ to the likelihood of $X = 1$ when $Y = 0$, the empirical version of the ratio of odds ratios is unchanged. Therefore, we can use $\alpha = \frac{ad}{bc}$ as a measure of the association between X and Y, with a value of α close to 1 indicating a low degree of association. Values of α far from 1, either close to 0 or very large, indicate a high degree of association.

Reciprocals of the cross-product ratio indicate the same degree of association, but in different direction. For example, values of α of 3 and 1/3 indicate the same degree of association. If $\alpha = 3$, $Y = 1$ is more likely when $X = 1$; if, $\alpha = 1/3$, $Y = 1$ is more likely when $X = 0$.

For the data in Table 5.1 on the relationship between starting quarterbacks who were a top 10 pick and the team making the playoffs, $\alpha = 0.6$. This means that the odds that a team with a quarterback who was a top 10 pick makes the playoffs are only 0.6 as large as the odds that a team with a quarterback who was not a top 10 pick makes the playoffs. Alternatively, the odds that a playoff team has a quarterback who was a top 10 pick are only 0.6 as large as the odds that a non-playoff team has quarterback who was a top 10 pick.

The main advantage of the cross-product ratio as a measure of association is that it is easy to interpret. However, the cross-product ratio does have the disadvantage that its range is 0 to ∞ with 1 indicating no relationship between the variables. Although there is nothing inherently wrong with these properties, it is sometimes easier to work with a measure that has properties similar to those of a correlation coefficient. Let

$$Q = \frac{\alpha - 1}{\alpha + 1} \tag{5.6}$$

The quantity Q, known as Yule's Q, has the same information as the cross-product ratio but now the range is -1 to 1, with 0 indicating no relationship. Furthermore, like a correlation coefficient, negative and positive values of Q with the same magnitude indicate the same degree of association, but in opposite direction. For example, for $\alpha = 3$, $Q = 1/2$; for $\alpha = 1/3$, $Q = -1/2$. On the other hand, Q does not have the same simple interpretation in terms of the relative odds ratios that α does.

For the data in Table 5.1 on the relationship between starting quarterbacks who were a top 10 pick and making the playoffs, $\alpha = 0.6$, so that $Q = -0.25$. It follows that quarterbacks who were top 10 picks are slightly less likely to make the playoffs, the same general conclusion we reached using the correlation coefficient, which is -0.119.

One advantage of α and Q over the correlation is that, unlikely the correlation, they are not sensitive to the row to the row and column totals. For instance, consider the example on the relationship between complete games and league. For the hypothetical data in Table 5.4, in which all the complete games are thrown by American League pitchers, $\alpha = \infty$ and therefore, $Q = 1$. That is, based on Q, there is perfect association between complete games and league, as would be expected using this data; recall that the value of the correlation for this hypothetical data is only 0.176. For the actual data, as given in Table 5.5, $Q = 0.15$.

In general, if the row totals and column totals are greatly different, as in Table 5.5, either α or Q is a more powerful measure of the association between the variables than is the correlation. If the row and column totals are roughly similar, as in Table 5.1 (where the row totals are 11 and 21 and the column totals are 12 and 20), then Q and r often show the same general level of association; α still has the advantage of being easier to interpret.

Even if there is actually no association between the variables, we know that the measure of association calculated from data, such as the correlation or Q, will not be exactly 0. Therefore, it is often of interest to determine if an observed association is statistically significantly from 0. First, note that if either the correlation of Q is 0, then the other measure is 0 as well. To see this, note that for the correlation to be 0, we must have

$$ad - bc = 0$$

If this holds, then $ad = bd$, so that $\alpha = 1$ and, therefore, $Q = 0$. Conversely, $Q = 0$ only if $\alpha = 1$, in which case $ad = bc$, so that the correlation is 0 as well. Therefore, to determine if an observed association is 0, we can compare the correlation r to $2/\sqrt{n}$, as we did when analyzing continuous data.

For instance, for the example on the relationship between starting quarterbacks who were a top 10 pick and making playoffs, $r = 0.119$ and $n = 32$, so that $2/\sqrt{n} = 0.35$; therefore, the observed association is not statistically significant.