Primer on Correlation Coefficients

Researchers are often interested in how two continuous variables relate to one another. To examine the relationship between body mass and fasting blood sugar, for example, one might study 20 people and measure both variables in each. The simplest approach to examine the relationship is to draw a picture, a scatterplot (an x–y graph), of body mass vs. fasting blood sugar. In this case, there are 20 dots, each representing one person. Scatterplots of other relationships may involve different units of analysis, as shown in Table 1.

Any of these relationships can also be quantified by a single number—the correlation coefficient, also known as *r*. Because journals frequently only publish the number (and not the picture), this primer offers three questions to help readers visualize and interpret correlation coefficients.

TABLE 1

VARIABLE 1	VARIABLE 2	UNIT OF ANALYSIS
Body mass	Fasting blood sugar	Individual
Pneumococcal vaccination compliance	Years in practice	Physician practice
Mammography compliance	Pap smear compliance	Clinic
Physicians per capita	Death rate	State

What Is the Sign on the Coefficient?

The first step is to look at the sign on *r*. If *r* is a positive number, the variables are directly related. In other words, as one goes up, so does the other (height and weight are a good example). If *r* is a negative number, the variables are inversely related. In other words, as one goes up, the other goes down (an example might be age and exercise capacity in adults). Knowing the sign helps you visualize the slope in the scatterplot, as shown in Figure 1.

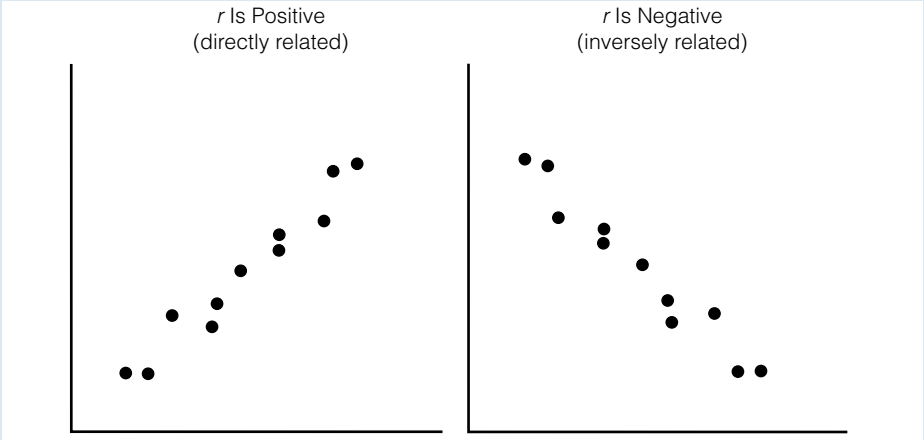

FIGURE 1.

What Is the Magnitude of the Coefficient?

The next step is to consider how big *r* is; *r* ranges from –1 to 1. An *r* of 0 signifies absolutely no correlation, whereas an *r* of –1 or 1 signifies a perfect correlation (all the data points fall on a line). In practice, *r* always has some intermediate value—there's always some correlation between two variables, but it's never perfect. The bigger the absolute value of *r* (i.e., the closer to –1 or 1), the stronger the correlation. The smaller the absolute value (i.e., the closer to 0), the weaker the correlation.

To provide perspective on what various *r*'s look like, Figure 2 shows three positive correlation coefficients and their associated scatterplots. (The scatterplots for the negative correlation coefficients would simply be mirror images.) Note that it may be difficult to see a relationship when *r* is less than 0.3 (or greater than –0.3).

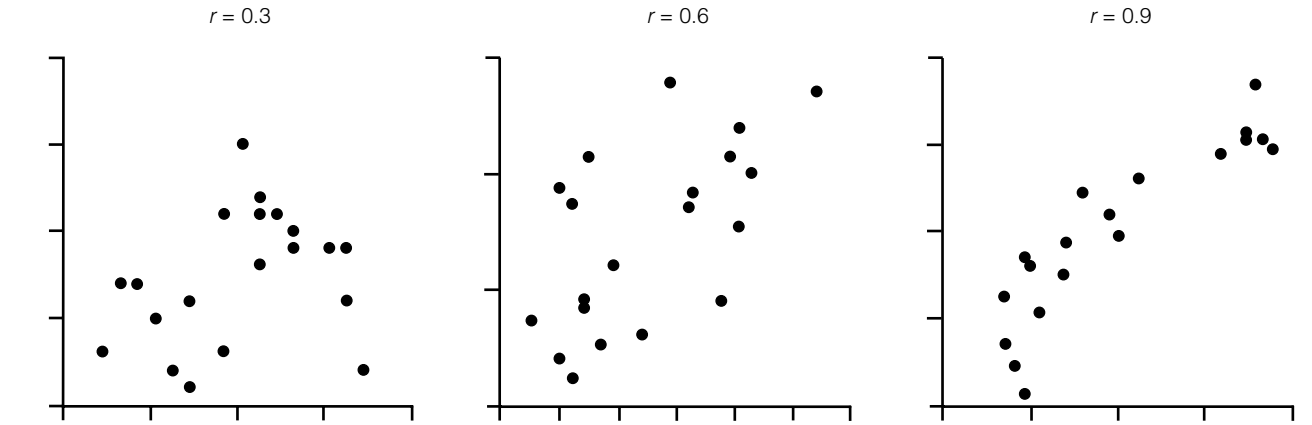

FIGURE 2.

The absolute magnitude of *r* is also a major determinant of statistical significance (the other being the number of observations). Consider 20 observations as depicted above. An *r* of 0.3 (a weak correlation) has an associated *P* value of 0.2. The *P* value falls with stronger correlations: *P* = 0.005 for an *r* of 0.6 and *P* < 0.0001 for an *r* of 0.9.

Does the Coefficient Reflect a General Relationship or an Outlier?

A critical reader will want to consider if seeing a scatterplot might influence the interpretation of *r*. As shown in Figure 3, a single extreme data point (an outlier) can have a powerful effect on the correlation coefficient when the sample size is small.

To mitigate this problem, *r* is often recalculated substituting ranks for the raw data. (An *r* calculated using raw data is called a Pearson *r*, while an *r* calculated using ranks is called a Spearman *r*. A reported *r* should be assumed to be Pearson *r* unless otherwise noted.)

For example, fasting blood sugar levels of 610, 320, 290, and 280 mg/dL would be converted to ranks 1, 2, 3, and 4; body weights of 350, 270, 220, and 210 lb would be converted to ranks 1, 2, 3, and 4; and the data point (610, 220) would become (1, 3). This recalculation does not eliminate the effect of outliers, but it does help to dampen their effects (in Figure 3, from left to right the recalculated *r*'s are 0.56, 0.62, and 0.37). In small samples, this recalculation can be particularly important.

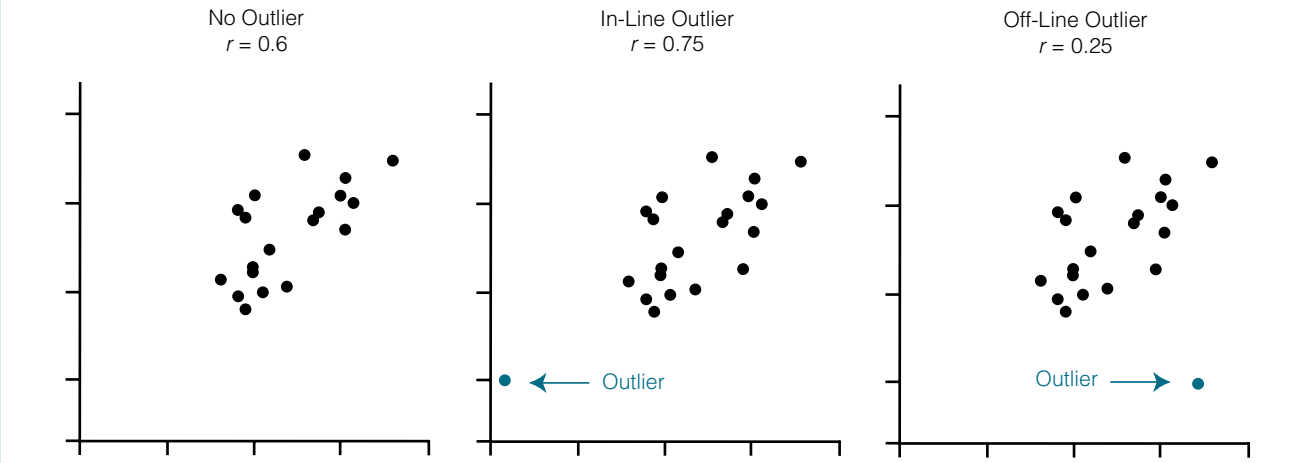

FIGURE 3.

Although correlation coefficients are an efficient way to communicate the relationship between two variables, they are not sufficient to interpret a relationship. The unit of analysis also matters. For example, a strong positive correlation between influenza and pneumococcal vaccination rates measured among physicians should be interpreted differently than the same coefficients measured among clinics. The former may imply that physicians have different beliefs about vaccinations, whereas the latter may simply reflect that clinics differ in the resources devoted to vaccination (e.g., reminder systems, nurse-run vaccination clinics).

Finally, correlation coefficients do not communicate information about whether one variable moves in response to another. There is no attempt to distinguish between the two variables—that is, to establish one as dependent and the other as independent. Thus, relationships identified using correlation coefficients should be interpreted for what they are: associations, not causal relationships.

A compendium of **ecp** *primers from past issues can be viewed and/or requested at http://www.acponline.org/journals/ecp/primers.htm.*