2

Fundamental Concepts

You should bring to a journey of learning about SEM prerequisite knowledge about some fundamental statistical concepts. One is the technique of multiple regression (MR). Although MR analyzes observed variables only, many of the principles that underlie it generalize directly to SEM. Next, the correct interpretation of statistical tests in general is considered, as are some special issues about their use in SEM. The basic logic of bootstrapping, a computer-based resampling procedure with increasing application in SEM, is also discussed. Some advice: *Even if you think that you already know some of these topics, you should nevertheless read this whole chapter carefully.* This is because many readers tell me that they learned something new after hearing about the issues outlined next.

MULTIPLE REGRESSION

I assume that you are already familiar with bivariate correlation and regression. You can find reviews of these topics in just about any introductory statistics book.[1] The logic of MR is considered next for the case of two continuous predictors, X_1 and X_2, and a continuous criterion Y, but the same ideas apply when there are ≥ 3 predictors. Pearson correlations among the predictors and the criterion are represented with the symbols r_{Y1}, r_{Y2}, and r_{12}. These coefficients are known as **zero-order correlations** because they do not control for intercorrelation. For example, r_{Y1} does not control for the possibility that $r_{Y2} \neq 0$ (X_2 also covaries with Y) or that $r_{12} \neq 0$ (the predictors are correlated). Features of MR especially relevant to SEM are emphasized next.

[1]See G. Garson's online *StatNotes* for a review: *http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm*

Ordinary Least Squares Estimation

With two predictors, the form of the unstandardized regression equation is

$$\hat{Y} = B_1 X_1 + B_2 X_2 + A \tag{2.1}$$

where $\hat{Y}$ is a predicted score. The term $\hat{Y}$ is a **composite**, or a weighted linear combination of the two predictors, X_1 and X_2. Equation 2.1 has both a covariance structure and a mean structure. The covariance structure corresponds to the unstandardized regression coefficients B_1 and B_2, and the mean structure to the intercept (constant) A. The values of B_1, B_2, and A are estimated with the method of **ordinary least squares** (OLS) so that the **least squares criterion** is satisfied. The latter means the sum of the squared residuals, or $\Sigma(Y - \hat{Y})^2$, is as small as possible in a particular sample. The method of OLS estimation is a **partial-information method** or a **limited-information method** because it analyzes the equation for only one criterion at a time.

Residuals in OLS estimation are uncorrelated with each of the predictors. That is,

$$r_{(Y-\hat{Y})1} = r_{(Y-\hat{Y})2} = 0 \tag{2.2}$$

where the residuals are represented in each subscript by the term $(Y - \hat{Y})$ and the predictors X_1 and X_2 by, respectively, the terms 1 and 2. The equality represented in Equation 2.2 is required in order for the computer to derive a unique set of regression weights that satisfies the least squares criterion. Conceptually, assuming the independence of residuals and predictors permits estimation of the relative predictor power of the latter (e.g., B_1 for X_1), with omitted (unmeasured) predictors held constant. Bollen (1989) refers to this assumption as **pseudoisolation** of the measured from the unmeasured predictors. Other implications of this assumption are considered later.

The overall multiple correlation between the predictors and the criterion, $R_{Y\cdot12}$, is actually just the Pearson correlation between the observed and predicted scores, or

$$R_{Y\cdot12} = r_{Y\hat{Y}} \tag{2.3}$$

Unlike Pearson correlations, though, the range of multiple correlations is 0–1.0. The value of $R^2_{Y\cdot12}$ indicates the proportion of explained variance. For example, if $R_{Y\cdot12} = .40$, then $R^2_{Y\cdot12} = .16$, so we can say that X_1 and X_2 together explain 16% of the total variance in Y. The values of B_1, B_2, and A in Equation 2.1 in a particular sample are those that maximize predictive power. Consequently, OLS estimation **capitalizes on chance**, which implies that (1) $R^2_{Y\cdot12}$ tends to overestimate the population proportion of explained variance ρ^2, and (2) it is possible that similar values of B_1, B_2, and A may not be found in a replication sample.

There are many corrections that downward adjust R^2 values as a function of sample size and the number of predictors (Yin & Fan, 2001). Perhaps the most common correction is Wherry's (1931) equation

$$\hat{R}^2 = 1 - (1 - R^2)\left(\frac{N-1}{N-k-1}\right) \tag{2.4}$$

where $\hat{R}^2$ is the adjusted estimate of ρ^2 and k is the number of predictors. The statistic $\hat{R}^2$ is a **shrinkage-corrected R^2**. In small samples, the value of $\hat{R}^2$ can be quite a bit less than that of R^2. The value of the former can even be negative; in this case, $\hat{R}^2$ is interpreted as though its value were zero. As the sample size increases for a constant number of predictors, values of $\hat{R}^2$ and R^2 are increasingly similar, and in very large samples they are essentially equal. That is, it is unnecessary to correct for positive bias in very large samples.

Regression Weights

The **unstandardized regression coefficients** B_1 and B_2 in Equation 2.1 indicate the expected raw score difference in Y, given a difference of a single point in one predictor while we are controlling for the other. For example, if $B_1 = 5.40$ and $B_2 = 3.65$, then the expected difference on Y is 5.40 points given a difference on X_1 of 1 point, with X_2 held constant. Likewise, a 1-point difference on X_2 predicts a 3.65-point difference on Y while controlling for X_1. Because unstandardized coefficients reflect the scales of their respective predictors, values of Bs from predictors with different raw score metrics are not directly comparable. Thus, one cannot conclude for this example that the relative predictive power of X_1 is greater than that of X_2 because $B_1 > B_2$. The **intercept** A is a constant that equals the value of $\hat{Y}$ when the scores on both predictors are zero ($X_1 = X_2 = 0$). It can be expressed as a function of the unstandardized coefficients and the means of all variables as follows:

$$A = M_Y - B_1 M_1 - B_2 M_2 \tag{2.5}$$

In contrast, means have no bearing on the values of the regression coefficients B_1 and B_2.

The regression equation for standardized variables is

$$\hat{z}_Y = b_1 z_1 + b_2 z_2 \tag{2.6}$$

where z_1 and z_2 are, respectively, standardized scores (normal deviates[2]) on X_1 and X_2, and b_1 and b_2 are, respectively, the **standardized regression coefficients**. The latter are also called **beta weights** because each standardized coefficient estimates a population parameter designated by the symbol β. Beta weights indicate the expected difference on the criterion in standard deviation units, controlling for all other predictors. Also, their values can be directly compared across predictors. For example, if $b_1 = .40$, it means that the difference in Y is expected to be .40 standard deviations large, given a difference

[2] $z_1 = (X_1 - M_1)/SD_1$, $z_2 = (X_2 - M_2)/SD_2$.

on X_1 of one full standard deviation controlling for X_2. The term b_2 has the analogous meaning except that X_1 is held constant. If $b_1 = .40$ and $b_2 = .80$, then we could say that the relative predictive power of X_2 is exactly twice that of X_1 in standard deviation units because the ratio b_2 / b_1 equals .80/.40, or 2.0.

Because beta weights are adjusted for intercorrelations among the predictors (and with the criterion, too), their absolute values are usually lower than those of the corresponding bivariate correlations (e.g., $b_1 = .40$, $r_{Y1} = .60$). This is not always true, though. Absolute values of b weights can exceed those of the corresponding correlation (e.g., $b_1 = .80$, $r_{Y1} = .60$). It is also possible for absolute values of b weights to exceed 1.0 or even for the signs of a beta weight and the corresponding correlation to be in opposite directions (e.g., $b_1 = -.40$, $r_{Y1} = .20$). When any of these cases occur, a suppression effect is indicated. Suppression is dealt with later.

For two predictors, the formulas for their beta weights are

$$b_1 = \frac{r_{Y1} - r_{Y2}\, r_{12}}{1 - r_{12}^2} \quad \text{and} \quad b_2 = \frac{r_{Y2} - r_{Y1}\, r_{12}}{1 - r_{12}^2} \tag{2.7}$$

The numerators in Equation 2.7 reflect one aspect of holding the other predictor constant.[3] In the formula for b_1, for example, the product of both bivariate correlations that involve the other predictor, X_2, is literally subtracted out of the bivariate correlation for X_1. The denominators in Equation 2.7 adjust the total standardized variance by removing the proportion shared by the two predictors. When there are ≥ 3 predictors, the formulas for the beta weights are more complicated but follow the same principles. The relation between unstandardized and standardized regression weights is expressed next:

$$B_1 = b_1 \left(\frac{SD_Y}{SD_1} \right) \quad \text{and} \quad B_2 = b_2 \left(\frac{SD_Y}{SD_2} \right) \tag{2.8}$$

The statistic $R^2_{Y \cdot 12}$ can also be expressed as a function of the beta weights and the bivariate correlations of the predictors with the criterion. With two predictors,

$$R^2_{Y \cdot 12} = b_1\, r_{Y1} + b_2\, r_{Y2} \tag{2.9}$$

The role of beta weights as corrections for the other predictor is apparent in this equation. Specifically, if $r_{12} = 0$ (the predictors are independent), then $b_1 = r_{Y1}$ and $b_2 = r_{Y2}$ (Equation 2.7), which means that $R^2_{Y \cdot 12}$ is just the sum of r^2_{Y1} and r^2_{Y2}. However, if $r_{12} \neq 0$ (the predictors covary), then b_1 and b_2 do not equal the corresponding bivariate correlations and $R^2_{Y \cdot 12}$ is not the simple sum of r^2_{Y1} and r^2_{Y2} (it is less).

As mentioned, beta weights can be directly compared across different predictors within the same sample. However, it is *not* generally correct to directly compare beta

[3] In a bivariate regression analysis with a single predictor X, the standardized regression coefficient is r_{XY}, the Pearson correlation with the criterion Y.

weights for the same predictors but across different samples, especially if those samples have different variances. This is because beta weights are standardized based on the variability in a particular sample (e.g., Equation 2.8 but solved for each of b_1 and b_2). If the within-group variances are not the same, then the basis of that standardization is not constant.[4] It is usually better to compare *unstandardized* regression coefficients across different samples. The same point holds in SEM analyses: It is the unstandardized solution that we directly compare across groups.

Presented in Table 2.1 is a small data set with scores on X_1, X_2, and Y. Assume that scores on these variables are from, respectively, a test of working memory, phonics skill, and reading achievement. Exercise 1 for this chapter will ask you to calculate and interpret the results for these data summarized next:

$$R_{Y \cdot 12} = .801, \quad R^2_{Y \cdot 12} = .641$$
$$B_1 = .242, \quad B_2 = .193, \quad A = 10.771$$
$$b_1 = .320, \quad b_2 = .599$$

An alternative to using a commercial computer program for the chapter exercises is a freely available calculating webpage for MR.[5] See the website for this book (p. 3) for links to other online calculating pages.

Assumptions

The statistical assumptions of MR are stringent, probably more so than many researchers realize. They are summarized next:

1. Regression weights reflect linear relations only. If there are also curvilinear relations, then values of regression weights will underestimate predictive power.
2. Statistical tests in MR assume that the residuals are normally distributed and have uniform variances across all levels of the predictors. The latter characteristic is **homoscedasticity**, and its opposite, **heteroscedasticity**, can be caused by outliers, severe non-normality in the observed scores, or more measurement error at some levels of the criterion or predictors. In the next chapter I will show you how to screen your data for heteroscedasticity.
3. It is assumed that the scores on the predictors are perfectly reliable (no measurement error). This assumption is necessary because there is no direct way in MR to represent less-than-perfect score reliability for the predictors. Consequences of minor

[4]Here is another example: Suppose that the same multiple-choice exam is administered in each of two different classes. For each class, scores are reported as the proportion correct, but relative to the highest score in each group, not the total number of items. Although the proportions in each class are standardized and have the same range (0–1.0), they are not directly comparable across the classes if the highest scores in each group are unequal.

[5]*http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/MultRgression.htm*

TABLE 2.1. Example Data Set for Multiple Regression

Case	X_1	X_2	Y
A	3	65	24
B	8	50	20
C	10	40	22
D	15	70	32
E	19	75	27

violations of this requirement may not be critical, but more serious ones can result in bias. This bias can affect not only the regression weights of predictors measured with error but also those of other predictors. However, it is difficult to anticipate the direction of this **error propagation**. Depending on sample intercorrelations, some regression weights may be biased upward (too large), but others may be biased in the other direction. There is no requirement that the criterion should be measured without error, but the use of a psychometrically inadequate measure of it can reduce the value of R^2. When the predictors are measured without error but the criterion is measured with error, beta weights tend to be too small, but not the unstandardized regression weights. If the predictors are measured with error, too, then these effects for the criterion could be amplified, diminished, or canceled out, but it is best not to hope for the latter. See Liu (1988) for more information.

4. It is assumed that omitted predictors are uncorrelated with measured predictors, or those in the equation. This requirement is a consequence of the fact that the residuals are uncorrelated with the predictors in OLS estimation. This is a strong assumption, one that is probably violated in most applications of MR (and SEM, too). This assumption also concerns the issue of specification error, which is considered next.

Specification Error

Specification error refers to the problem of omitted predictors that account for some unique proportion of total criterion variance but are not included in the analysis. A related term is **left-out-variable error** or, more lightheartedly, the "heartbreak of L.O.V.E." The idea of specification error in SEM is even broader than in MR, but the omission of relevant predictors is a concern in SEM, too. Suppose that $r_{Y1} = .40$ and $r_{Y2} = .60$ for, respectively, predictors X_1 and X_2. A researcher measures only X_1 and uses it as the sole predictor of Y. The standardized regression coefficient for the *included predictor* in this bivariate analysis is $r_{Y1} = .40$. If the researcher had the foresight to also measure X_2, the *omitted predictor*, and enter it along with X_1 as a predictor in an MR analysis, the beta weight for X_1 in this analysis may not equal .40. If not, then r_{Y1} as a standardized regression coefficient with X_1 as the sole predictor does not reflect the true predictive power of X_1 compared with b_1 derived with both predictors in the equation. However, the difference between r_{Y1} and b_1 varies with r_{12}, the correlation between the included and omitted predictors. Specifically, if the included and omitted predictors are unrelated

($r_{12} = 0$), there is no difference ($r_{Y1} = b_1$) because there is no correction for correlated predictors. But as the absolute value of their correlation increases ($r_{12} \neq 0$), the amount of the difference between r_{Y1} and b_1 due to the omission of X_2 becomes greater.

Presented in Table 2.2 are the results of three pairs of regression analyses. In all pairs, X_2 is considered the omitted predictor.[6] One member of each pair of analyses is a bivariate regression with X_1 as the sole predictor, and the other member is an MR with both X_1 and X_2 in the equation. Constant across all three sets of analyses are the bivariate correlations between the predictors and the criterion ($r_{Y1} = .40$, $r_{Y2} = .60$). The only thing that varies across the three sets is the value of r_{12}, the correlation between the predictors. Reported for each analysis in Table 2.2 are the standardized regression weights (r_{Y1} for the bivariate regression; b_1 and b_2 for the MR) and also the overall multiple correlation (R_{Y12}) for the regression of Y on both X_1 and X_2. For each case in the table, compare in the same row the value of r_{Y1} in boldface with that of b_1, also in boldface. The difference between these values (if any) indicates the amount by which the bivariate standardized regression coefficient for X_1 does not accurately reflect its predictive power relative to when X_2 is also in the equation.

Note in Table 2.2 that when the omitted predictor X_2 is uncorrelated with the included predictor X_1 (case 1, $r_{12} = 0$), the standardized regression weight for X_1 is the same regardless of whether or not X_2 is in the equation ($r_{Y1} = b_1 = .40$). However, when $r_{12} = .30$ (case 2), the value of b_1 is lower than that of r_{Y1}, respectively, .24 versus .40. This happens because b_1 controls for the correlation between X_1 and X_2, whereas r_{Y1} does not. Thus, r_{Y1} *overestimates* the association between X_1 and Y relative to b_1. In case 3 in the table, the correlation between the included and omitted predictors is even higher ($r_{12} = .60$), which for these data results in an even greater discrepancy between r_{Y1} and b_1 (respectively, .40 vs. .06).

Omitting a predictor correlated with others in the equation does not always result in overestimation of the predictive power of an included predictor. For example, if X_1 is the included predictor and X_2 is the omitted predictor, it is also possible for the absolute value of r_{Y1} to be *less* than that of b_1 (i.e., r_{Y1} *underestimates* the relation indicated by b_1)

TABLE 2.2. Examples of the Omitted Variable Problem

	Predictor(s)			
		Both X_1 and X_2		
Case	X_1 only	X_1	X_2	$R_{Y \cdot 12}$
1. $r_{12} = 0$	.40	.40	.60	.72
2. $r_{12} = .30$	.40	.24	.53	.64
3. $r_{12} = .60$	.40	.06	.56	.60

Note. Numerical values for X_1 and X_2 are standardized regression coefficients. For all cases, X_2 is considered the omitted variable; $r_{Y1} = .40$ and $r_{Y2} = .60$.

[6]The same principles hold if X_1 is the omitted predictor and X_2 is the included predictor.

or even for r_{Y1} and b_1 to have different signs. Both cases indicate suppression. However, overestimation due to omission of a predictor probably occurs more often than underestimation (suppression). Also, the pattern of bias may be more complicated when there are several included and omitted variables (e.g., overestimation for some included predictors; underestimation for others).

Predictors are typically excluded because they are not measured. Thus, it is difficult to know by how much and in what direction regression coefficients may be biased relative to what their values would be if all relevant predictors were included. However, it is unrealistic to expect the researcher to know and be able to measure all relevant predictors. In this way, all regression equations are probably misspecified to some degree. If omitted predictors are uncorrelated with included predictors, the consequences of specification error may be slight. Otherwise, the consequences may be more serious. Careful review of theory and research is the main way to avoid a serious specification error by decreasing the potential number of left-out variables.

Suppression

Perhaps the most general definition is that **suppression** occurs when either the absolute value of a predictor's beta weight is greater than its bivariate (zero-order) correlation with the criterion or the two have different signs. So defined, suppression implies that the estimated relation between a predictor and a criterion while controlling for the other predictors is a "surprise," given the bivariate correlations. Suppose that X_1 is amount of psychotherapy, X_2 is degree of depression, and Y is number of prior suicide attempts. The bivariate correlations in a hypothetical sample are

$$r_{Y1} = .19, \quad r_{Y2} = .49, \quad \text{and } r_{12} = .70$$

Based on these results, it may seem that psychotherapy is harmful because of its positive association with suicide attempts ($r_{Y1} = .19$). When both predictors (depression, psychotherapy) are entered as predictors in the same regression equation, however, the results are

$$b_1 = -.30, \quad b_2 = .70, \quad \text{and } R_{Y\cdot12} = .54$$

The beta weight for psychotherapy (–.30) has the opposite sign of its bivariate correlation with the criterion (.19), and the beta weight for depression (.70) exceeds its bivariate correlation (.49).

The "surprising" results just described are due to controlling for other predictors. Here, people who are more depressed are also more likely to be in psychotherapy ($r_{12} = .70$). Depressed people are more likely to try to harm themselves ($r_{Y2} = .49$). Corrections for these associations in MR reveal that the relation of psychotherapy to suicide attempts is actually negative once depression is controlled. It is also true that the relation of depression to suicide attempts is even stronger once psychotherapy is

controlled. Omit either psychotherapy or depression from the analysis—a specification error—and the bivariate regression results with the remaining predictor are misleading. This example concerns **negative suppression**, where the predictors have positive correlations with the criterion and each other, but one receives a negative beta weight in the analysis.

A second type of suppression is **classical suppression**, where one predictor is uncorrelated with the criterion but receives a nonzero beta weight controlling for another predictor. For example, given the following correlations in a hypothetical sample,

$$r_{Y1} = 0, \quad r_{Y2} = .60, \quad \text{and } r_{12} = .50$$

the results of an MR analysis are

$$b_1 = -.40, \quad b_2 = .80, \quad R_{Y\cdot12} = .69$$

This example of classical suppression (i.e., $r_{Y1} = 0$, $b_1 = -.40$) demonstrates that bivariate correlations of zero can mask true predictive relations once other variables are controlled. There is also **reciprocal suppression**, which can occur when two predictors correlate positively with the criterion but negatively with each other. See Shieh (2006) for more information about suppression.

Death to Stepwise Regression, Think for Yourself

There are two basic ways to enter predictors into the equation: One is to enter all predictors at once, or **simultaneous entry**. The other is to enter predictors over a series of steps, or **sequential entry**. Entry order can be determined according to one of two different standards, theoretical (rational) or empirical (statistical). The rational standard corresponds to **hierarchical regression**, where *you* tell the computer a fixed order of entry for the predictors. For example, sometimes demographic variables are entered at the first step, and then entered at the second step is a psychological variable of interest. This order not only controls for the demographic variables but also permits evaluation of the predictive power of the psychological variable, over and beyond that of simple demographic variables.

An example of the statistical standard is **stepwise regression**, where the *computer* selects predictors for entry based on statistical significance (e.g., which predictor, if entered into the equation, would have the most statistically significant regression weight?). After they are selected, predictors at a later step can also be removed from the equation according to statistical test outcomes (e.g., if a predictor's regression weight is no longer statistically significant). The stepwise process stops when there could be no statistically significant increase in R^2 by adding more predictors. There are variations on stepwise regression—for example, some methods select predictors but do not later remove them (**forward inclusion**), and others begin with all predictors in the equation and then automatically remove them (**backward elimination**)—but all such methods are directed by the computer, not you.

Stepwise regression and related methods pose many problems, so many that such methods are now basically forbidden in some research areas (e.g., Thompson, 1995), and for good reason, too. One problem is extreme capitalization on chance. Another is that not all regression computer programs print correct values of statistical tests in stepwise regression; that is, the computer's choices may actually be wrong. Both of these problems imply that whatever final set of predictors happen to be selected by the computer in empirically driven procedures is unlikely to replicate. Worst of all, such methods give the illusion that the researcher does not have to think about the problem. Sribney (1998) offers this advice: "Personally, I would no more let an automatic routine select my model than I would let some best-fit procedure pack my suitcase" (Ronan Conroy's Comments section, para. 8).

In SEM, there are methods for modifying structural equation models with poor fit to the data that are analogous to empirically based methods in MR. These methods in SEM indicate the particular effects that would result in the greatest improvement in fit if those effects were added to the model. Some SEM computer tools, such as LISREL, offer an **automatic modification** (AM) option that mechanically adds effects according to statistical criteria. Such purely exploratory options greatly capitalize on chance; they also give the illusion that you need not think about the problem. I do not recommend the use of AM-type options. Instead, the modification of your model should be guided mainly by your hypotheses, just as its specification in the first place should be so guided. There is a role in SEM for more limited empirically based methods, but they should be used in a way that respects your hypotheses. These issues are elaborated in Chapter 8, on hypothesis testing in SEM.

PARTIAL CORRELATION AND PART CORRELATION

The technique of **partial correlation** concerns the phenomenon of **spuriousness**: if the observed relation between two variables is due to ≥ 1 common cause(s), their association is spurious. To illustrate this concept, consider these zero-order correlations between vocabulary breadth (Y), shoe size (X_1), and age (X_2) in a hypothetical sample of children not all the same age:

$$r_{Y1} = .50, \quad r_{Y2} = .60, \quad \text{and } r_{12} = .80$$

Although the correlation between shoe size and vocabulary breadth is fairly substantial (.50), it is hardly surprising because both are caused by a third variable, age (i.e., maturation).

The partial correlation $r_{Y1\cdot2}$ removes the influence of a third variable X_2 from both X_1 and Y. The formula is

$$r_{Y1\cdot2} = \frac{r_{Y1} - r_{Y2}\, r_{12}}{\sqrt{(1 - r_{Y2}^2)(1 - r_{12}^2)}} \tag{2.10}$$

The denominator in Equation 2.10 adjusts the total standardized variance of both Y and X_1 for their overlap with X_2. Applied to the hypothetical correlations just listed, the partial correlation between shoe size and vocabulary breadth controlling for age is $r_{Y1\cdot 2}$ = .04. (An exercise will ask you to calculate this partial correlation.) Because the association between X_1 and Y essentially disappears when X_2 is controlled, their observed relation r_{Y1} = .50 may be a spurious one. The technique of SEM readily allows the representation of presumed spurious associations due to common causes.

Equation 2.10 for partial correlation can be extended to control for two or more external variables. For example, the higher-order partial correlation $r_{Y1\cdot 23}$ estimates the association between X_1 and Y controlling for both X_2 and X_3. There is a related coefficient called **part correlation** or **semipartial correlation** that partials external variables out of either of two variables, but not both. The formula for the part correlation $r_{Y(1\cdot 2)}$ for which the association between X_1 and X_2 is controlled but not the association between Y and X_2 is presented next:

$$r_{Y(1\cdot 2)} = \frac{r_{Y1} - r_{Y2}\, r_{12}}{\sqrt{1 - r_{12}^2}} \tag{2.11}$$

Note that the denominator in Equation 2.11 adjusts the total standardized variance only for the overlap of X_1 with X_2. Given the same bivariate correlations among these three variables reported earlier, the part correlation between vocabulary breadth (Y) and shoe size (X_1) controlling only the latter for age (X_2) is $r_{Y(1\cdot 2)}$ = .03. This result (.03) is somewhat smaller than the partial correlation for these data, or $r_{Y1\cdot 2}$ = .04. In general, $r_{Y1\cdot 2}$ is larger in absolute value than $r_{Y(1\cdot 2)}$. An exception is when r_{12} = 0; in this case, $r_{Y1\cdot 2}$ = $r_{Y(1\cdot 2)}$.

Relations among the squares of the various correlations just described can be nicely illustrated with a Venn-type diagram like the one in Figure 2.1. The circles represent the total standardized variances of the criterion Y and the predictors X_1 and X_2. The regions in the figure labeled *a*–*d* make up the total standardized variance of Y, so

$$a + b + c + d = 1.0$$

Areas *a* and *c* in the figure represent the portions of Y uniquely predicted by, respectively, X_1 and X_2, but area *b* represents the simultaneous overlap (redundancy) of the predictors with Y. Area *d* represents the proportion of unexplained variance. The squared zero-order correlations of the predictors with the criterion and the overall squared multiple correlation can be expressed as sums of the areas *a*, *b*, *c*, or *d* in Figure 2.1, as follows:

$$r_{Y1}^2 = a + b \quad \text{and} \quad r_{Y2}^2 = b + c$$
$$R_{Y\cdot 12}^2 = a + b + c = 1.0 - d$$

The squared part correlations correspond directly to the unique areas *a* and *c* in Figure 2.1. Each of these areas also equals the *increase* in the total proportion of explained variance that occurs by adding a second predictor to the equation. That is,

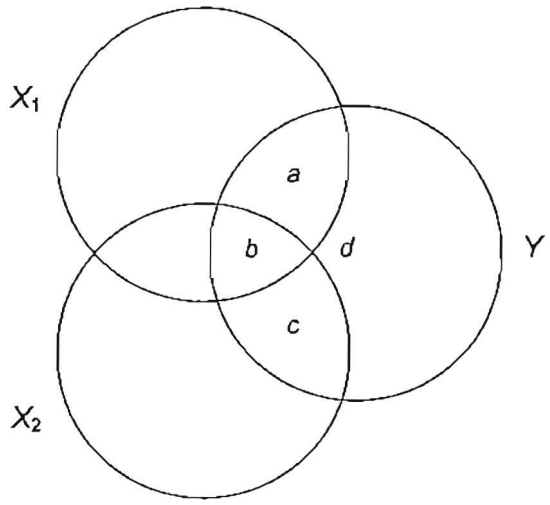

FIGURE 2.1. Venn diagram for the standardized variances of Y, X_1, and X_2.

$$r^2_{Y(1\cdot2)} = a = R^2_{Y\cdot12} - r^2_{Y2} \qquad (2.12)$$
$$r^2_{Y(2\cdot1)} = c = R^2_{Y\cdot12} - r^2_{Y1}$$

In contrast, the squared partial correlations correspond to areas a, c, and d in Figure 2.1, and each estimates the proportion of variance in the criterion explained by one predictor but not the other. The formulas are

$$r^2_{Y1\cdot2} = \frac{a}{a+d} = \frac{R^2_{Y\cdot12} - r^2_{Y2}}{1 - r^2_{Y2}} \qquad (2.13)$$
$$r^2_{Y2\cdot1} = \frac{c}{c+d} = \frac{R^2_{Y\cdot12} - r^2_{Y1}}{1 - r^2_{Y1}}$$

Note that the numerator of each expression in Equation 2.13 is a squared part correlation. The denominators in Equation 2.13 correct the total standardized variance of the criterion for its overlap with the other predictor. These denominators are generally < 1.0, which explains why squared partial correlations are generally larger than squared part correlations. Suppose that $R^2_{Y\cdot12} = .40$ and $r^2_{Y2} = .25$. These results follow:

$$r^2_{Y(1\cdot2)} = .40 - .25 = .15$$
$$r^2_{Y1\cdot2} = .15/(1 - .25) = .20$$

In words, predictor X_1 uniquely explains .15, or 15% of the total variance of Y (squared part correlation). Of the variance in Y not already explained by X_2, predictor X_1 accounts

for .20, or 20% of the remaining variance (squared partial correlation). See G. Garson (2009) for an online review of partial correlation and part correlation.[7]

When predictors are correlated—which is just about always—beta weights, partial correlations, and part correlations are alternative ways to describe in standardized terms the relative explanatory power of each predictor controlling for the rest. None is more "correct" than the other because each gives a different perspective on the same data. However, remember that unstandardized regression coefficients (*B*) are preferred when comparing results for the same predictors across different samples.

OTHER BIVARIATE CORRELATIONS

When all observed variables are continuous, it is Pearson correlations that are usually analyzed in SEM as part of analyzing covariances. (Recall that cov_{XY} is the product of r_{XY} and the standard deviations of each variable; Equation 1.1.) However, noncontinuous variables can be analyzed in SEM, too, so you need to know something about other kinds of bivariate correlations. There are other forms of the Pearson correlation for observed variables that are either categorical or ordinal. For example:

1. The **point-biserial correlation** (r_{pb}) is a special case of *r* that estimates the association between a dichotomous variable and a continuous one (e.g., gender, weight).
2. The **phi coefficient** ($\hat{\varphi}$) is a special case for two dichotomous variables (e.g., treatment-control, relapsed-not relapsed).
3. **Spearman's rank order correlation** or **Spearman's rho** is for two ranked variables.

It is also possible in SEM to analyze non-Pearson correlations that assume the underlying data (i.e., on a latent variable) are continuous and normally distributed instead of discrete. For example:

1. The **biserial correlation** is for a continuous variable and a dichotomy (e.g., agree-disagree), and it estimates what the Pearson *r* would be if both variables were continuous and normally distributed.
2. The **polyserial correlation** is the generalization of the biserial correlation that does basically the same thing for a continuous variable and a categorical variable with three or more levels.
3. The **tetrachoric correlation** for two dichotomous variables estimates what *r* would be if both variables were continuous and normally distributed.

[7] *http://faculty.chass.ncsu.edu/garson/PA765/partialr.htm*

4. The **polychoric correlation** is the generalization of the tetrachoric correlation that estimates *r* but for categorical variables with two or more levels.

Computing polyserial or polychoric correlations is complicated (Nunnally & Bernstein, 1994) and requires specialized software such as PRELIS, which is the part of LISREL for manipulating, generating, and transforming data. The PRELIS program can be used to estimate polyserial or polychoric correlations, depending on the types of variables in the data set. It can also estimate results for **censored variables**, which have large proportions of their scores at minimum or maximum values. Consider the variable "price paid for a new car in the last year." In a hypothetical sample, only 10% bought a new car year in the last year, so the scores for rest (90%) are zero. This variable is censored because not everyone buys a new car every year. Instead of deleting the 90% of the cases who did not purchase a new car, PRELIS would attempt to estimate results for this variable in the whole sample assuming that the underlying distribution is normal. Options for analyzing non-Pearson correlations in SEM are considered in Chapter 7.

LOGISTIC REGRESSION

Sometimes outcome variables are dichotomous or binary variables. Examples include graduated–did not graduate and survived–died. Some options to analyze dichotomous outcomes in SEM are based on the logic of **logistic regression** (LR). This technique is generally used instead of MR when the criterion is dichotomous. Just as in MR, the predictors in LR can be either continuous or categorical. However, the regression equation in LR is a logistic function that approximates a nonlinear relation between the dichotomous outcome and a linear combination of the predictors. An example of a logistic function for a hypothetical sample is illustrated in Figure 2.2. The closed circles in the figure represent along the *Y*-axis whether cases with the same illness either improved ($Y = 1.0$) or did not improve ($Y = 0$). Along the *X*-axis, the closed circles in the figure represent scores on a composite variable made up of various indexes of healthy behavior (exercise, preventative care, etc.). The logistic function fitted to the data in Figure 2.2 is S-shaped, or **sigmoidal** in form. This function generates predicted probabilities of improvement, given scores on the healthy behavior composite.

The estimation method in logistic regression is not OLS. Instead, it is usually ML estimation but is applied after transforming the binary outcome into a **logit variable**, which is typically the natural logarithm—base *e*, or approximately 2.71828—of the **odds** of the target outcome. The latter tell us how much more likely it is that a case is a member of the target group instead of a member of the other group (Wright, 1995), and it equals the probability of the target outcome divided by the probability of the other outcome. An example follows.

Suppose that 60% of the cases improved over a particular time, but the rest, or 40%, did not. Assuming that improvement is the target outcome, the odds of improvement are calculated here as .60/.40, or 1.5. That is, the odds are 3:2 in favor of improvement.

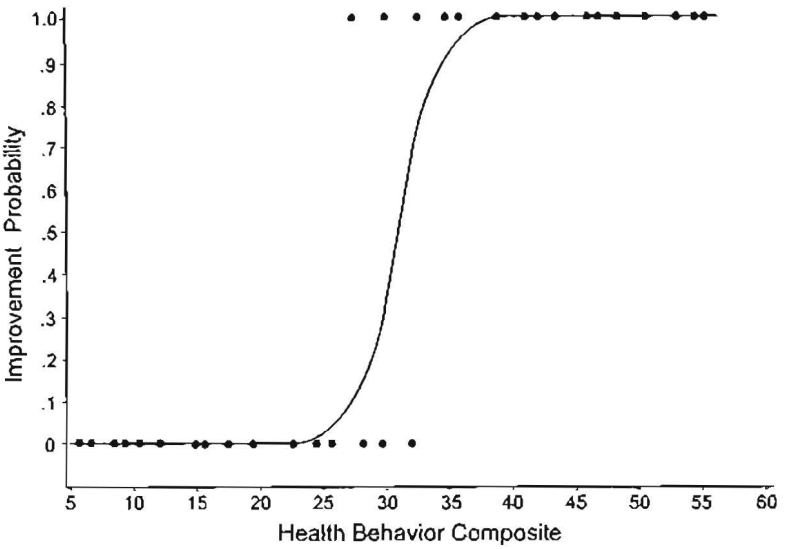

FIGURE 2.2. Example of a logistic function where closed circles represent actual data values and the curve represents predicted probabilities.

Regression coefficients for each predictor in LR can be converted into an **odds ratio**, which estimates the difference in the odds of the target outcome for a one-point difference in the predictor, controlling for all other predictors. For example, if the estimated odds ratio for amount of exercise were 5.60, then the odds of improvement are 5.6 times greater for each one-point increase on the exercise variable, holding constant other predictors. Values of odds ratios less than 1.0 would indicate for this example a relative reduction in the odds of improvement given higher scores on that predictor, and odds ratios that equal 1.0 would indicate no difference in improvement odds for any value of the predictor. See Peng, Lee, and Ingersoll (2002) for more information about LR.

STATISTICAL TESTS

Characteristics of statistical tests especially relevant for SEM are emphasized next.

Standard Errors

Perhaps the most basic form of a statistical test is the **critical ratio**, which is the ratio of a sample statistic over its **standard error**. The standard error is the standard deviation of a **sampling distribution**, which is a probability distribution of a statistic based on all possible random samples, each based on the same number of cases. A standard error estimates **sampling error**, the difference between sample statistics and the corresponding population parameter. Given constant variability among population cases, standard

error varies inversely with sample size. This means that distributions of statistics from larger samples are generally narrower (less variable) than distributions of the same statistic from smaller samples.

There are textbook formulas for the standard errors of statistics with simple distributions. By "simple" I mean that (1) the statistic estimates a single parameter and (2) the shape of the distribution is not a function of that parameter. For example, the textbook formula for estimating the standard error of the mean is

$$SE_M = \frac{SD}{\sqrt{N}} \tag{2.14}$$

It is more difficult to estimate standard errors for statistics that do not have simple distributions. There are approximate methods amenable to hand calculation for some statistics, such as sample proportions, where distribution shape and variability depend on the value of the population proportion. Such methods generate **asymptotic standard errors** that assume a large sample. However, if your sample is not large, such estimated standard errors may not be accurate. But some other statistics, such as the multiple correlation R, have distributions so complex that there may be no approximate standard error formula for hand calculation. Estimation of standard error in such cases may require specialized software (Kline, 2004, chap. 4). In SEM, standard errors for effects of latent variables are estimated by the computer, but these estimates are just that. This means that their values could change if, say, a different estimation method is used. So do not overinterpret results of statistical tests for latent variables.

Power and Types of Null Hypotheses

In large samples under the assumption of normality, a critical ratio is interpreted as a z-statistic in a normal curve with a mean of zero and a standard deviation that equals the standard error. A rule of thumb for large samples is that if the absolute value of this z-statistic exceeds 2.00, the null hypothesis (H_0) that the corresponding parameter is zero is rejected at the .05 level of statistical significance ($p < .05$) for a two-tailed test (H_1). The precise value of z for the .05 level is 1.96 and for the .01 level it is 2.58. Within small samples, critical ratios approximate a t-distribution instead of a z-distribution, which necessitates the use of special tables to determine critical values of t for the .05 or .01 levels. Within large samples, t and z for the same sample statistic are essentially equal.

The failure to reject some null hypothesis is a meaningful outcome only if (1) the power of the test is adequate and (2) the null hypothesis is at least plausible to some degree. Briefly, **power** is the probability of rejecting the null hypothesis when there is a real effect in the population (H_1 is true, H_0 is not). Power varies directly with the magnitude of the real population effect and your sample size. Other factors that affect power include:

1. The level of statistical significance (e.g., .05 vs. .01) and the directionality of H_1 (i.e., one- or two-tailed tests).

2. Whether the samples are independent or dependent (i.e., a between- or within-subject design).
3. The particular test statistic used.
4. The reliability of the scores.

The following combination generally leads to the greatest power: a large sample, the .05 level of statistical significance, a one-tailed (directional) H_1, a within-subject design, a parametric test statistic (e.g., t) rather than a nonparametric statistic (e.g., Mann–Whitney U), and scores that are very reliable. The power of a study should be estimated when the study is planned but *before* the data are collected (Wilkinson & the Task Force on Statistical Inference, 1999). Ideally, power should be as high as possible, such as > .85. If power is only about .50, then the odds of rejecting a false null hypothesis are no greater than guessing the outcome of a coin toss. In fact, tossing a coin instead of conducting the study would be just as likely to give the correct decision and would save time and money, too (Schmidt & Hunter, 1997). How to estimate power in SEM is described in a later chapter, but the typical power of certain kinds of statistical tests in SEM are often relatively low even in large samples.

The type of null hypothesis tested most often in the behavioral sciences is a **nil hypothesis**, which says that the value of a population parameter or the difference between two parameters is zero. A nil hypothesis for the t-test of a mean contrast is

$$H_0\colon \mu_1 - \mu_2 = 0$$

(i.e., $H_0\colon \mu_1 = \mu_2$), which predicts that two population means are exactly equal. However, it is unlikely that the value of *any* population parameter (or difference between two parameters) is exactly zero, especially if zero implies the complete absence of an effect or association. It is also possible to specify a **non-nil hypothesis** for the t-test, such as

$$H_0\colon \mu_1 - \mu_2 = 5.00$$

but this is rarely done in practice. As the name suggests, a non-nil hypothesis is a statement that a population difference or effect is not zero.

It is more difficult to specify and test non-nil hypotheses for other types of statistical tests, such as the F-test when comparing ≥ 3 means. This is because computer programs almost always assume a nil hypothesis. Nil hypotheses may be appropriate when it is unknown whether effects exist at all, such as in new research areas where studies are mostly exploratory. Such hypotheses are less suitable in established research areas when it is already known that an effect is probably not zero. Perhaps most statistical results reported in literature are associated with nil hypotheses that are implausible. An example of an implausible nil hypothesis in the environmental sciences is the assumption of equal survival probabilities for juvenile and adult members of a species (Anderson, Burnham, & Thompson, 2000). When a nil hypothesis is implausible, then (1) it is

a "straw man" argument (a fallacy) that is easily rejected and (2) estimated probabilities of data (p) under that unlikely hypothesis are too low.

It is important not to misinterpret the outcome of a statistical test in any type of data analysis. See Topic Box 2.1 for a review of the "Big Five" misinterpretations of statistical significance.

Statistical Tests in SEM

Here is a critical point about statistical tests in SEM: In ML estimation (and in some other methods, too), *standard errors are generally calculated for the unstandardized solution only.* You can see this fact when you look through the output of an SEM computer tool and find no standard errors printed for standardized estimates. This means that

TOPIC BOX 2.1

The "Big Five" Misinterpretations of Statistical Significance*

There is ample evidence that many of us do not know the correct interpretation of outcomes of statistical tests, or p values. For example, at the end of a standard statistics course, most students know how to calculate statistical tests, but they do not typically understand what the results mean (Haller & Krauss, 2002). About 80% of psychology professors endorse at least one incorrect interpretation of statistical tests (Oakes, 1986). It is easy to find similar misinterpretations in books and articles (Cohen, 1994), so it seems that psychology students get their false beliefs from teachers and also from what students read. However, the situation is no better in other behavioral science disciplines (e.g., Hubbard & Armstrong, 2006).

Most misunderstandings about statistical tests involve overinterpretation, or the tendency to see too much meaning in statistical significance. Specifically, we tend to believe that statistical tests tell us what we want to know, but this is wishful thinking. Elsewhere I described statistical tests as a kind of collective Rorschach inkblot test for the behavioral sciences in that what we see in them has more to do with fantasy than with what is really there (Kline, 2004). Such wishful thinking is so pervasive that one could argue that much of our practice of hypothesis testing based on statistical tests is myth.

In order to better understand misinterpretations of p values, let us first deal with their correct meaning. Here it helps to adopt a **frequentist perspective** where probability is seen as the likelihood of an outcome over repeatable events under constant conditions except for chance (sampling error). From this view, a probability does not apply directly to a single, discrete event. Instead, probabil-

*Part of this presentation is based on Kline (2009, chap. 5).

ity is based on the expected relative frequency over a large number of trials, or in the long run. Also, there is no probability associated with whether or not a particular guess is correct in a frequentist perspective. The following mental exercises illustrate this point:

1. A die is thrown, and the outcome is a 2. What is the probability that this particular result is due to chance? The correct answer is *not* $p = 1/6$, or .17. This is because the probability .17 applies only in the long run to repeated throws of the die. In this case, we expect that .17 of the outcomes will be a 2. The probability that any particular outcome of the roll of a die is the result of chance is actually $p = 1.00$.
2. One person thinks of a number from 1 to 10. A second person guesses that number by saying, 6. What is the probability that the second person guessed right? The correct answer is *not* $p = 1/10$, or .10. This is because the particular guess of 6 is either correct or incorrect, so no probability (other than 0 for "wrong" or 1.00 for "right") is associated with it. The probability .10 applies only in the long run after many repetitions of this game. That is, the second person should be correct about 10% of the time over all trials.

Let us now review the correct interpretation of statistical significance. You should know that the abbreviation *p* actually stands for the conditional relative-frequency probability:

$$p\left(\begin{array}{c|c} \text{Result or} & H_0 \text{ true, random sampling,} \\ \text{more extreme} & \text{other assumptions} \end{array}\right)$$

which is the likelihood of a sample result or one even more extreme (a range of results) assuming that the null hypothesis is true, the sampling method is random sampling, and all other assumptions for the corresponding test statistic, such as the normality requirement of the *t*-test, are tenable. Two correct interpretations for the specific case $p < .05$ are given next. Other correct definitions are probably just variations of the ones that follow:

1. Assuming that H_0 is true (i.e., every result happens by chance) and the study is repeated many times by drawing random samples from the same population, less than 5% of these results will be even more inconsistent with H_0 than the particular result observed in the researcher's sample.
2. Less than 5% of test statistics from random samples are further away from the mean of the sampling distribution under H_0 than the one for the observed result. That is, the odds are less than 1 to 19 of getting a result from a random sample even more extreme than the observed one.

Described next are what I refer to as the "Big Five" false beliefs about *p* values.

Three of the beliefs concern misinterpretation of p, but two concern misinterpretations of their complements, or $1 - p$. Approximate base rates for some of these beliefs, reported by Oakes (1986) and Haller and Krauss (2002) in samples of psychology students and professors, are reported beginning in the next paragraph. What I believe is the biggest of the Big Five is the **odds-against-chance fallacy**, or the false belief that p indicates the probability that a result happened by chance (e.g., if $p < .05$, then the likelihood that the result is due to chance is < 5%). Remember that p is estimated for a range of results, not for any particular result. Also, p is calculated assuming that H_0 is true, so the probability that chance explains any individual result is already taken to be 1.0. Thus, it is illogical to view p as somehow measuring the probability of chance. I am not aware of an estimate of the base rate of the odds-against-chance fallacy, but I think that it is nearly universal in the behavioral sciences. It would be terrific if some statistical technique could estimate the probability that a particular result is due to chance, but there is no such thing.

The **local Type I error fallacy** for the case $p < .05$ is expressed as follows: I just rejected H_0 at the .05 level. Therefore, the likelihood that this particular (local) decision is wrong (a Type I error) is < 5% (70% approximate base rate among psychology students and professors). This belief is false because any particular decision to reject H_0 is either correct or incorrect, so no probability (other than 0 or 1.00; i.e., right or wrong) is associated with it. It is only with sufficient replication that we could determine whether or not the decision to reject H_0 in a particular study was correct. The **inverse probability fallacy** goes like this: Given $p < .05$; therefore, the likelihood that the null hypothesis is true is < 5% (30% approximate base rate). This error stems from forgetting that p values are probabilities of data under H_0, not the other way around. It would be nice to know the probability that either the null hypothesis or alternative hypothesis were true, but there is no statistical technique that can do so based on a single result.

Two of the Big Five concern $1 - p$. One is the **replicability fallacy**, which for the case of $p < .05$ says that the probability of finding the same result in a replication sample exceeds .95 (40% approximate base rate). If this fallacy were true, knowing the probability of replication would be useful. Unfortunately, a p value is just the probability of the data in a particular sample under a specific null hypothesis. In general, replication is a matter of experimental design and whether some effect actually exists in the population. It is thus an empirical question and one that cannot be directly addressed by statistical tests in a particular study. Here I should mention Killeen's (2005) p_{rep} statistic, which is a mathematical transformation of $1 - p$ (i.e., generally, $p_{rep} \neq 1 - p$) that *estimates* the average probability of getting a result of the same sign (direction) in

a hypothetical replication, assuming random sampling. Killeen suggested that p_{rep} may be less subject to misinterpretation than p values, but not everyone agrees (e.g., Cumming, 2005). It is better to actually conduct replication studies than rely on statistical prediction.

The last of the Big Five, the **validity fallacy**, refers to the false belief that the probability that H_1 is true is greater than .95, given $p < .05$ (50% approximate base rate). The complement of p, or $1 - p$, is also a probability, but it is just the probability of getting a result even *less* extreme under H_0 than the one actually found. Again, p refers to the probability of the data, not to that of any particular hypothesis, H_0 or H_1. See Kline (2004, chap. 3) or Kline (2009, chap. 5) for descriptions of additional false beliefs about statistical significance.

It is pertinent to consider one last myth about statistical tests, and it is the view that the .05 and .01 levels of statistical significance, or α, are somehow universal or objective "golden rules" that apply across all studies and research areas. It is true that these levels of α are the conventional standards used today. They are generally attributed to Carl Fisher, but he did *not* advocate that these values be applied across all studies (e.g., Fisher, 1956). There are ways in decision theory to empirically determine the optimal level of α given estimate of the costs of various types of decision errors (Type I vs. Type II error), but these methods are almost never used in the behavioral sciences. Instead, most of us automatically use $\alpha = .05$ or $\alpha = .01$ without acknowledging that these particular levels are arbitrary. Even worse, some of us may embrace the **sanctification fallacy**, which refers to dichotomous thinking about p values that are actually continuous. If $\alpha = .05$, for example, then a result where $p = .049$ versus one where $p = .051$ is practically identical in terms of statistical outcomes. However, we usually make a big deal about the first (it's significant!) but ignore the second. (Or worse, we interpret it as a "trend" as though it was really "trying" to be significant, but fell just short.) This type of black-and-white thinking is out of proportion to continuous changes in p values. There are other areas in SEM where we commit the sanctification fallacy, and these will be considered in Chapter 8. This thought from the astronomer Carl Sagan (1996) is apropos: "When we are self-indulgent and uncritical, when we confuse hopes and facts, we slide into pseudoscience and superstition" (p. 27). Let there be no superstition between us concerning statistical significance going forward from this point.

results of statistical tests are available only for unstandardized estimates. Researchers often assume that results of statistical tests of unstandardized estimates apply to the corresponding standardized estimates. For samples that are large and representative, this assumption may not be problematic. You should realize, however, that the level of

statistical significance for an unstandardized estimate does not automatically apply to its standardized counterpart. This is true in part because standardized estimates have their own standard errors, and the ratio of a standardized statistic over its standard error may not correspond to the same p value as the ratio of that statistic's unstandardized counterpart over its standard error. This is why you should (1) always report the unstandardized estimates with their standard errors and (2) not associate results of statistical tests for unstandardized estimates with the corresponding standardized estimates. An example follows.

Suppose in ML estimation that the values of an unstandardized estimate, its standard error, and the standardized estimate are, respectively, 4.20, 2.00, and .60. In a large sample, the unstandardized estimate would be statistically significant at the .05 level because $z = 4.20/2.00$, or 2.10, which exceeds the critical value (1.96) at $p < .05$. Whether the standardized estimate of .60 is also statistically significant at $p < .05$ is unknown because it has no standard error. Consequently, it would be inappropriate to report the standardized estimate by itself as

× .60*

where the asterisk designates $p < .05$. It is better to report both the unstandardized and standardized estimates and also the standard error of the former, like this

✓ 4.20* (2.10) .60

where the standard error is given in parentheses and the asterisk is associated with the unstandardized estimate (4.20), not the unstandardized one (.60). Special methods in SEM for estimating correct standard errors for the standardized solution are described in Chapter 7.

Central and Noncentral Test Distributions

Conventional tests of statistical significance are based on central test distributions. A **central test distribution** assumes that the null hypothesis is true, and tables of critical values for distributions such as t, F, and χ^2 found in many introductory statistics textbooks are based on central test distributions. In a **noncentral test distribution**, however, the null hypothesis is *not* assumed to be true. Some perspective is in order. Families of central test distributions of t, F, and χ^2 are special cases of noncentral distributions of each test statistic just mentioned. Compared to central distributions, noncentral distributions have an extra parameter called the **noncentrality parameter**, which is often represented in the quantitative literature by the symbol Δ for the t statistic and by λ for the F and χ^2 statistics. This extra parameter indicates the degree of departure from the null hypothesis. An example follows.

Central t-distributions are described by a single parameter, the degrees of freedom df, but noncentral t-distributions are described by both df and Δ. Presented in Figure 2.3

are two t-distributions each where $df = 10$. For the central t-distribution in the left part of the figure, $\Delta = 0$. However, $\Delta = 4.17$ for the noncentral t-distribution in the right side of the figure. Note that the latter distribution in Figure 2.3 is positively skewed. The same thing happens but in the opposite direction for negative values of Δ for t-distributions. In a two-sample design, the positive skew in the noncentral t-distribution would arise due to sampling of positive mean differences because $\mu_1 - \mu_2 > 0$ (i.e., H_0: $\mu_1 - \mu_2 = 0$ is false).

Noncentral test distributions play an important role in certain types of statistical analyses. Computer programs that estimate power as a function of study characteristics and the expected population effect size analyze noncentral test distributions. This is because the concept of power assumes that the null hypothesis is false, and in a power analysis it is false to the degree indicated by the hypothesized effect size. A nonzero effect size generally corresponds to a value of the noncentrality parameter that is also not zero. Another application is the estimation of confidence intervals based on sample statistics that measure effect size, such as standardized mean differences (d) for mean contrasts or R^2 in regression analyses. Effect size estimation also assumes that the null hypothesis—especially when it is a nil hypothesis—is false. See Kline (2004) for more information about confidence intervals for effect sizes.

In SEM, some measures of model fit are based on noncentral test distributions, especially noncentral χ^2-distributions. These statistics indicate the degree of **approxi-**

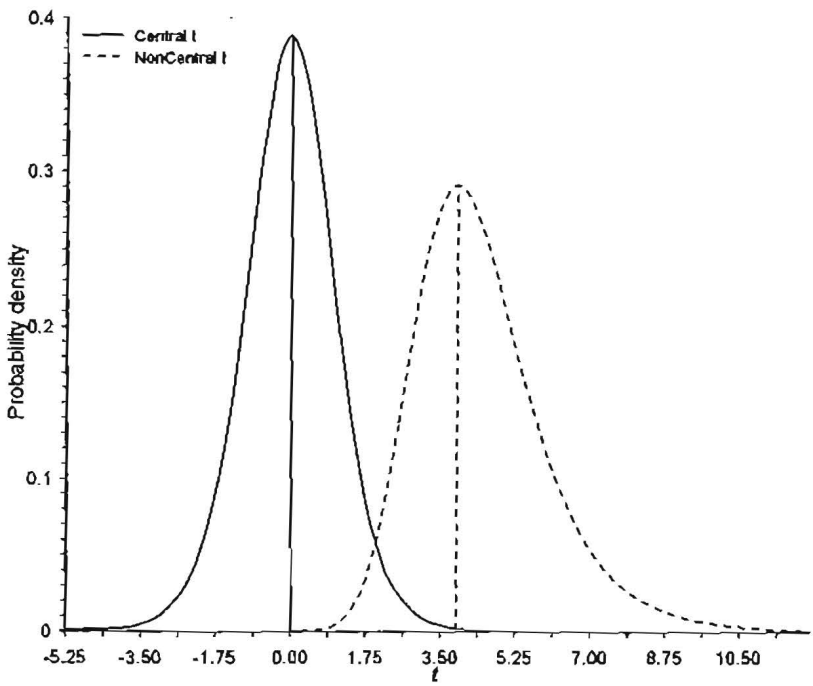

FIGURE 2.3. Distributions of central t and noncentral t for 10 degrees of freedom and where the noncentrality parameter equals 4.17 for noncentral t.

mate fit of your model to the data. That is, these fit indexes allow for an "acceptable" amount of departure from **exact fit** or **perfect fit** between model and data. What is considered "acceptable" departure from perfection is related to the estimated value of the noncentrality parameter for the χ^2 statistic that the computer calculates for your model. Other fit statistics in SEM measure the degree of departure from perfect fit, and these indexes are generally described by central χ^2-distributions. Assessment of model fit against these two standards, exact versus approximation, is covered in Chapter 8.

BOOTSTRAPPING

Bootstrapping is a computer-based method of resampling developed by B. Efron (e.g., 1979). There are two general kinds of bootstrapping. In **nonparametric bootstrapping**, your sample (i.e., data file) is treated as a pseudopopulation. Cases from the original data set are randomly selected with replacement to generate other data sets, usually with the same number of cases as the original. Because of sampling with replacement, (1) the same case can appear in more than one generated data set and (2) the composition of cases will vary slightly across the generated samples. When repeated many times (e.g., 1,000), bootstrapping simulates the drawing of numerous random samples from a population. Standard errors are estimated in this method as the standard deviation in the empirical sampling distribution of the same statistic across all generated samples. Nonparametric bootstrapping generally assumes only that the sample distribution has the same shape as that of the population distribution. In contrast, the distributional assumptions of many standard statistical tests, such as the *t*-test for means, are more demanding (e.g., normal and equally variable population distributions). A raw data file is necessary for nonparametric bootstrapping. This is not true in **parametric bootstrapping**, where the computer randomly samples from a theoretical probability density function specified by the researcher. This is a kind of Monte Carlo method that is used in computer simulation studies of the properties of particular estimators, including those of many used in SEM that measure model fit.

It is important to realize that bootstrapping is not a magical technique that can somehow compensate for small or unrepresentative samples, severely non-normal distributions, or the absence of actual replication samples. In fact, bootstrapping can potentially magnify the effects of unusual features in a small data set (Rodgers, 1999). More and more SEM computer programs, including Amos, EQS, LISREL, and Mplus, feature optional bootstrap methods. Some of these methods can be used to estimate the standard errors of a particular model parameter estimate or a fit statistic; bootstrapping can be used to calculate confidence intervals for these statistics, too. Bootstrapping methods are also applied in SEM to estimate standard errors for non-normal or categorical data and when there are missing data.

An example of the use of nonparametric bootstrapping to empirically estimate the standard error of a Pearson correlation follows. Presented in Table 2.3 is a small data set for two continuous variables where $N = 20$ and the observed correlation is $r_{XY} = .3566$.

TABLE 2.3. Example Data Set for Nonparametric Bootstrapping

Case	X	Y	Case	X	Y
A	12	16	K	16	37
B	12	46	L	13	51
C	21	66	M	18	32
D	16	70	N	12	53
E	18	27	O	22	52
F	16	27	P	12	34
G	16	44	Q	22	54
H	14	69	R	12	5
I	16	22	S	14	38
J	18	61	T	14	38

I used the nonparametric bootstrap procedure of SimStat for Windows (Version 2.5.5; Provalis Research, 1995–2004[8]) to resample from the data set in Table 2.3 in order to generate a total of 1,000 bootstrapped samples each with 20 cases. Presented in Figure 2.4 is the empirical sampling distribution of r_{XY} across the 1,000 bootstrapped samples. SimStat reported that the mean of this distribution is .3482 and the standard deviation is .1861. The former result (.3482) is close to the observed correlation (.3566), and the latter (.1861) is actually the bootstrapped estimate of the standard error of the observed correlation. The 95% bootstrapped confidence interval calculated by SimStat based on the distribution in the figure is –.0402 to .6490, and the bias-adjusted confidence interval is –.0402 to .6358.[9] One could use the method of nonparametric bootstrapping to estimate standard errors or confidence intervals for multiple correlations, too.

SUMMARY

Reviewed in this chapter were fundamental statistical concepts that underlie many aspects of SEM. One of these is the idea of statistical control—the partialing out of variables from other variables, a standard feature of most models in SEM. A related idea is that of spuriousness, which happens when an observed association between two variables disappears when controlling for common causes. The phenomenon of suppression is also related to statistical control. Suppression occurs in some cases when the sign of the adjusted relation between two variables differs from that of their bivariate correlation. One lesson of suppression is that values of observed correlations can mask true relations between variables once intercorrelations with other variables are controlled. Another is the importance of including all relevant predictors in the analysis. This is because the omission of predictors that are correlated with those included in the model is a specification error that may bias the results. Special issues concerning statistical

[8]You can download a free 30-day trial version of the full version of SimStat from *www.provalisresearch.com*

[9]In nonparametric bootstrapping, bias correction controls for lack of dependence due to possible selection of the same case ≥ 2 times in the same generated sample.

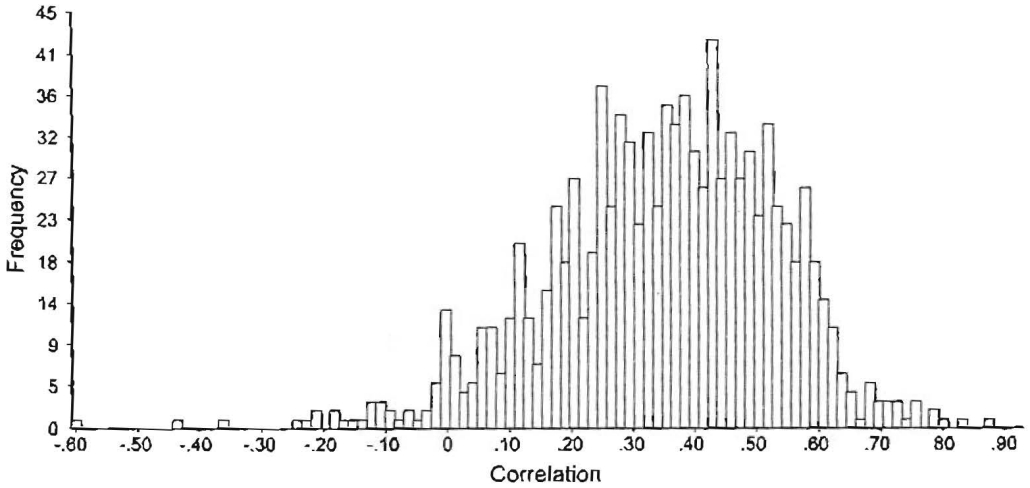

FIGURE 2.4. Empirical sampling distribution for the Pearson correlation r_{XY} in 1,000 bootstrapped samples for the data in Table 2.3.

tests were also considered, including the need to avoid common misinterpretations of statistical significance. Results of statistical tests in SEM generally apply to unstandardized estimates only, not to the corresponding standardized estimates. Also reviewed in this chapter was the basic logic of bootstrapping, a resampling technique that can be applied to estimate standard errors for statistics with complex distributions.

RECOMMENDED READINGS

The book by Cohen, Cohen, West, and Aiken (2003) is considered by many as a kind of "bible" of the multiple regression technique. The suggested chapters in Kline (2004) concern the correct interpretation of statistical tests and related statistics, such as standard errors (chaps. 1–3), and an introduction to bootstrapping (chap. 9). A more comprehensive review of bootstrap methods for estimation, regression, forecasting, and simulation is available in Chernick (2008).

Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). Hoboken, NJ: Wiley.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

EXERCISES

1. For the data in Table 2.1, calculate R_{Y12}, the unstandardized regression equation, and the standardized regression weights. Interpret the results assuming X_1, X_2, and Y are, respectively, measures of working memory, phonics skill, and reading achievement.

2. Calculate scores on $\hat{Y}$ and $Y - \hat{Y}$ for the data in Table 2.1. Show that $R_{Y12} = r_{Y\hat{Y}}$. Also show that the equality expressed in Equation 2.2 is true.

3. Calculate scores on $\hat{z}_Y$ and $z_Y - \hat{z}_Y$ for the data in Table 2.1. Show that the residuals in standardized form are uncorrelated with each predictor in standardized form.

4. Calculate a shrinkage-corrected R^2_{Y12} for the data in Table 2.1. Interpret the results.

5. Calculate R_{Y12}, b_1, and b_2, given $r_{Y1} = .40$, $r_{Y2} = .50$, and $r_{12} = -.30$. Describe the results.

6. Use Equation 2.10 to calculate the partial correlation between X and Y controlling for W, given these Pearson correlations: $r_{XY} = .50$, $r_{XW} = .80$, and $r_{YW} = .60$.

7. Suppose that the 95% confidence interval for the difference between two means in a particular sample is 75.25–84.60. Explain what is wrong with this statement: "There is a 95% chance that the interval 75.25–84.60 contains the population mean difference $\mu_1 - \mu_2$."

8. Find three incorrect definitions of statistical significance on the Internet. *Hint:* In Google, type "define: statistical significance." Explain what is wrong with each.