28

Sample Size Considerations

Research is what I'm doing when I don't know what I'm doing

- WERNHER VON BRAUN

INTRODUCTION

A common question when designing an RET is "what sample size should I use?" A related question is "can I analyze the data using statistical method X given my sample size?" The answers to these questions are complicated. A major facet of statistical inference that sample size affects is sampling error. Sample size also affects the properties of estimators and, in cases of asymptotic theory, our ability to calculate coherent margins of errors and p values. I begin this chapter by reviewing key concepts of sampling error, properties of estimators, and asymptotic theory to set the foundation for my discussion of sample size decision making. With this as background, I then discuss power analysis per se, the choice

of sample size to minimize margins of error, the conduct of localized simulations to choose sample sizes, and analytic methods that can be used with small sample sizes. The chapter is long, so you probably will need to process it in several sittings.

When making statistical conclusions associated with null hypothesis testing, there are two types of errors you can make. You can reject the null hypothesis when you should not have (i.e., you falsely say there is a relationship between variables when, in fact, there isn't), or you can fail to reject the null hypothesis when you should have (i.e., you miss a relationship that exists). The first type of error is called a **Type I error** and the second type is called a **Type II error**. The probability of making a Type II error is called **beta** and power is one minus beta. **Power** is the probability of correctly rejecting the null hypothesis or the probability of *not* making a Type II error. It reflects effect sensitivity. Power of 0.90 means that we are 90% likely to detect an effect that is present in the population; power of 0.80 means that we are 80% likely to detect an effect in the population; power of 0.70 means we are 70% likely to detect an effect; and so on.

What is an acceptable level of statistical power? Tradition in the social sciences is to seek power of 0.80 or greater, but this is an arbitrary standard. You need to think more deeply about such matters and decide what you believe is a reasonable level of power to have. How bad would it be to miss a medication that non-trivially reduces the death rate from cervical cancer? How bad would it be to miss the presence of a serious side effect to a new contraceptive method? I think we need to be more nuanced about setting acceptable levels of statistical power for a study rather than just blindly accepting a 0.80 standard.

Most researchers know that sample size impacts statistical power with larger sample sizes being associated with increased power. However, sample size decisions are complex and require us to think about more than statistical power. We also need to think about margins of errors, asymptotic theory, covariance matrix stability, model complexity, effects on estimation properties, missing data, and practical constraints. The present chapter considers these facets of sample size decision making.

SAMPLING ERROR

Consider the case of 1,000 families in a small town. An investigator wants to describe the average number of children in the families and decides to interview heads of households to determine family sizes. Because of practical limitations, the investigator is unable to include all 1,000 families in the study, so s/he instead resorts to a (random) sample. Assume that the sample size is ten. In this case, the population is the 1,000 families in the town and the sample is the ten families who are selected to be interviewed. Suppose the true population mean is 3.50, the true variance is 2.09, and the true standard deviation is 1.45

but the investigator is unaware of these values.

For the random sample of 10 families, the number of children in each family is found to be 3, 4, 4, 5, 2, 4, 1, 1, 4, and 3, which yields a mean of 3.10. Note that this value is not equal to the true population mean. A sample statistic can (and usually does) differ from the value of its corresponding population parameter because of sampling error. Sampling error occurs because the sample estimate is based on only a portion of the overall population. The amount of sampling error is represented as the difference between the value of a sample statistic and the value of the corresponding population parameter. In our example, the amount of sampling error is 3.10 - 3.50 = -.40. In practice, an investigator does not know the value of the population parameter, so it is impossible to know the exact amount of sampling error. Sampling error occurs in most research and is something we must deal with when we make inferences about populations.

We run the risk of making Type II errors when sampling error is large. This is because with large amounts of sampling error, there is more "noise" in our data that can lead us astray. If we can harness sampling error, we usually can lower the risk of Type II errors which is key to increasing statistical power. As such, it is important to understand factors that contribute to sampling error and how we can bring sampling error under some control.

Factors Affecting Sampling Error

There are multiple factors that impact sampling error, but I concentrate on two of them. One factor is sample size. In general, the larger the sample size, the less the sampling error, everything else being equal. In our family size example, if I sample only 10 families, I likely will have more sampling error than if I sample, say, 999 families. The second factor is the variability of scores in the population. I can illustrate this principle using a simplistic example. Consider two populations each with 5 observations and the following scores:

Population A	Population B
2	4
3	4
4	4
5	4
6	4

The mean in each population is 4, but the two populations obviously differ in the variability of scores. Suppose I do not know the value of the means and I am told I can randomly sample two cases from each population to estimate it. In Population A, I might end up sampling the scores 4 and 6 when I select the two observations randomly, and the average

of them is 5. Absent any other information, 5 is my best guess about the population mean and, as it turns out, I am "off" by 1 unit (sampling error). For Population B, I randomly select the scores of, say, the third and fifth person, which are 4 and 4 and the mean is 4. There is no sampling error. Because there is no variability in scores in Population B, it does not matter which two cases I happen to sample because they all equal 4 and, when averaged, will yield the value of the Population B mean. In Population A, where the scores are variable, there are many combinations of two that I could sample, some of which will yield means that are quite discrepant from the population mean.

Other factors can impact sampling error depending on the parameter but more often than not these factors can be traced back to sample size and variability. As examples, for proportions, the closer a population proportion is to 0.50, the more sampling error there will be, everything else being equal. This is because for a binary variable, there is more variability in scores when the proportion is 0.50 (half the scores are 0s and half are 1s) than when the proportion is, say, 0.90 (10% of the scores are zeros and 90% are 1s) or 0.10 (90% of the scores are zeros and 10% are 1s). For correlations, the closer a population correlation is to zero, the more sampling error there will be, everything else being equal. This is because when the true population correlation is, say 0.90, there is not as much variability in the multivariate cloud of scores than when the correlation is zero, everything else being equal.

Sample size and population variability usually are key in statistical theory as influencers of sampling error. If I have small sample sizes, there likely will be more sampling error; if the variables I am studying have considerable population variability, either univariately or multivariately, there likely will be more sampling error.

Sampling Distributions and Standard Errors

In practice, we can never know how much sampling error is operating in a study because we never know the true value of the population parameter we seek to estimate. However, we can use our knowledge of factors that impact sampling error to estimate or get a sense of how much sampling error might be operating in a given study. To make this intuitive, I need to introduce an abstract concept called a **sampling distribution**.

Suppose I am interested in characterizing the annual starting salary of new assistant professors in the United States but, because of practical constraints, I must use a random sample of 100 professors to estimate this (N = 100). In theory, there are many different combinations of 100 individuals I could end up with in this study. One random sample of 100 individuals might yield a sample mean of $45,132. Another random sample of 100 individuals might yield a sample mean of $48,215. Yet another random sample of 100 individuals might yield a sample mean of $50,108. None of these sample means are

probably exactly equal to the true population mean because of sampling error, but I hope they all are at least close to the population mean.

Now think about the thousands of means that would result if I were to take every possible random sample of size 100 relative to the population of assistant professors and calculated a mean for each one of them. Of course, we would never do this in practice, but in theory, it is possible. The resulting set of means that I would generate from this exercise is called a **sampling distribution of the mean**. Note that I can compute a sampling distribution for any statistic, not just the mean. For a correlation between X and Y in a population, I could calculate a correlation in every possible random sample of a given size and I would end up with thousands of correlations. These correlations represent a **sampling distribution of correlations**.

One would hope that the sampling distribution of a statistic, in this case the mean, has very little variability in it. Stated another way, it should make us nervous if we know there are wild fluctuations in the parameter estimate as we move from one random sample to the next; after all, when we conduct a study, we are focusing on one sample in the larger sampling distribution. An index of variability of sample means in a sampling distribution of the mean is the standard deviation of the sampling distribution (if we could, in fact, calculate it – but we can't because sampling distributions are hypothetical). A standard deviation of zero would mean that every sample yielded the exact same result. By contrast, larger standard deviations indicate more sample-to-sample fluctuations. The standard deviation of a sampling distribution is called the **standard error**. One can speak of the standard error of a mean, the standard error of a correlation, the standard error of a regression coefficient – i.e., the standard error of any parameter that we seek to estimate.

Although we can rarely calculate a standard error, it turns out statisticians have devised methods for estimating standard errors from sample data. The methods need not concern us here and formulas that statisticians use can be found in statistical texts. My emphasis is more on helping you appreciate what a standard error is than actually computing an estimate of it. Statistical software will do that for you.

The estimated standard error of the mean reflects the accuracy with which sample means tend to estimate a population mean. If I am trying to estimate the average number of children that couples have in their completed families, and I tell you that the estimated standard error of the mean for samples of size 50 is 1.30, this means that, on average, the sample means differ 1.30 units (in this case, "children") from the true population mean. This reflects a sizable amount of error. If, on the other hand, the estimated standard error of the mean is only 0.10, this would mean that, on average, sample means deviate only 0.10 from the true population mean.

If you examine the formula for estimating a standard error of the mean, you will find

that the two factors I discussed earlier as determinants of sampling error are part of the formula – the sample size and the population standard deviation (which is reflected by the variability of scores in the sample or the sample standard deviation):

Estimated standard error = $SD / \sqrt{N}$

Note that in this formula as the sample size increases, the estimated standard error of the mean becomes smaller, other things being equal. Similarly, as the sample standard deviation (SD, which is an estimate of the population standard deviation) becomes smaller, so does the estimated standard error of the mean.

Keep in mind that interpretation of the size of a standard error is dependent on the metric of the variable(s) being studied. A standard error of 2.0 is quite large when estimating the average number of children people have in their families. However, a standard error of 2 would be amazingly small if one is estimating the average annual salary of assistant professors in the United States (i.e., on average, our sample means are only off by $2.). Computer output routinely reports estimated standard errors. They play a key role in significance tests and the formation of margins of error.

There is one other feature of sampling distributions I should mention. If you have a population with mean μ and standard deviation σ and you take sufficiently large random samples from the population when you calculate a mean, the sampling distribution of the mean will be approximately normally distributed even if the original variables themselves are not normally distributed. This property derives from a theorem known as the **central limit theorem** and is a property that statisticians take advantage of when estimating p values and confidence intervals.

SAMPLE SIZE AND PROPERTIES OF ESTIMATORS

Statisticians have elucidated properties of estimators, such as the sample mean as an estimator of the population mean or the sample correlation as an estimator of the population correlation. I highlight here three properties and then note their relevance to sample size decisions.

First is the property of **bias**. A good estimator is one that is not biased, everything else being equal. In statistics, bias has a precise meaning. An estimator is unbiased if the mean of the sampling distribution of the estimator equals the true value of the population parameter. A **positively biased estimator** is one whose sampling distribution mean is larger than the population parameter and a **negatively biased estimator** is one whose sampling distribution mean is smaller than the population parameter. The sample mean is an unbiased estimator of the population mean. A sample regression coefficient in a linear regression is an unbiased estimator of its corresponding population regression coefficient

if the assumptions of linear regression are met. Numerous statistics you are familiar with, however, are biased estimators. For example, the standard deviation of sample data is a positively biased estimator of the population standard deviation.[1] The sample squared correlation coefficient is a positively biased estimator of the population squared correlation coefficient.

Note that just because a sample estimator is unbiased does not mean that it is accurate. There will still be sampling error in the estimate for any given sample and it may or may not be close to the value of the true population parameter. All it means when we say an estimator is unbiased is that, on average, across all possible random samples of a given size, the mean of the estimate will equal the population parameter.

A second important property for estimators is that of **efficiency**. Efficiency refers to the sample-to-sample fluctuations of the estimator. Efficient estimators have low standard errors. Actually, efficiency is a relative term because it usually compares two estimators in terms of their relative efficiency. For example, for small sample sizes, eta squared is a more efficient estimator of the proportion of explained variance in the population compared to epsilon squared or omega squared.

A third important property of an estimator is its **consistency**. A consistent estimator is one whose probability of accurately reflecting the parameter in question increases with increasing sample size. Although this principle is a re-statement of my previous assertion that sample size influences sampling error, it turns out that some estimators lack such consistency; increasing N will not necessarily make the estimator more accurate. Inconsistent estimators are troubling.

When we choose estimators of population central tendencies, population variability, and population relationships between variables, we tend to prefer estimators that are unbiased, efficient, and consistent. To be sure, there are other properties that we take into account, such as the extent to which the estimator is outlier resistant. However, bias, efficiency, and consistency are three qualities we seek. Sometimes the properties of estimators are immune to sample size. For example, the mean of randomly selected independent scores from a population is an unbiased estimator of the population mean irrespective of sample size. By contrast, sometimes the properties of estimators are impacted by sample size; for example, the amount of bias in a sample squared multiple correlation as an estimator of the population squared multiple correlation is impacted by sample size. As such, we want to take the potential impact of sample size on the properties of estimators into account when choosing an N.

[1] This is why the formulas for variances, upon which standard deviations are based, divide the sum of squares by N-1 instead of N. Using N-1 is a correction factor that adjusts for the positive bias that results from using N.

SAMPLE SIZE AND ASYMPTOTIC THEORY

Asymptotic theory is a statistical framework for assessing the properties of estimators and statistical tests. A tenet of asymptotic theory is that sample size, in principle, can increase indefinitely. This tenet allows statisticians to evaluate the statistical properties of an estimator or a test of significance under the limit that $N \rightarrow \infty$, or as $f(N)$ goes to infinity. For example, statisticians characterize the limit of the function 1/N as 0 because as N gets larger and larger, the value of 1/N gets closer and closer to 0. It turns out that by invoking asymptotic theory, statisticians can derive many statistical principles that can't be derived when samples are finite. Invoking asymptotic theory is compatible with randomized trials because, as discussed in Chapter 4, we often conceptualize populations in such trials as hypothetical and extremely large. It also turns out that statistical properties derived using asymptotic theory often can be imported to the case of finite populations as long as the finite populations are sufficiently large. The question of what we mean when we say "sufficiently large" depends on the broader statistical context, as I discuss below.

Many of the statistical methods you are familiar with are based in asymptotic theory. This includes maximum likelihood SEM, logistic regression, chi square testing, and count/discrete regression, among others. These statistical tests usually are well behaved *as long as the sample size is large enough*. Again, the practical question becomes "what is large enough?" Sample size decisions also must take into account the demands of asymptotic theory if such theory underlies the tests/estimators that you use.

SAMPLE SIZE, COVARIANCE PROPERTIES, AND MODEL COMPLEXITY

When working with multivariate models, such as multiple regression, factor analysis, or multi-equation full information structural equation modeling (FISEM), another factor that must be weighed when making sample size decisions is the properties of the covariance/correlation matrix being analyzed in conjunction with the nature and complexity of the model being tested. Consider a multiple regression analysis with 3 predictors, X1, X2 and X3. Suppose the variables (including the outcome Y) all have a mean of zero and a standard deviation of 1.0 so that the input covariance matrix can be interpreted much like a correlation matrix. Consider the case where the true population covariance and correlation matrices between the variables has the following values:

	Y	X1	X2	X3
Y	1.00			
X1	0.40	1.00		
X2	0.35	0.50	1.00	
X3	0.30	0.50	0.50	1.00

The population regression equation for both the unstandardized and standardized versions of the equation is

Y = 0 + 0.275 X1 + 0.175 X2+ 0.075 X3

with X1 having the strongest coefficient. Suppose I select a random sample from this population and my sample size is 125. The values in the sample covariance and correlation matrices will not exactly equal the values of their population counterparts because of sampling error. Here is a correlation matrix I might observe in my sample:

	Y	X1	X2	X3
Y	1.00			
X1	0.41	1.00		
X2	0.34	0.55	1.00	
X3	0.32	0.51	0.44	1.00

The sample standardized regression equation in this case is

Y = 0.282 X1 + 0.142 X2+ 0.136 X3

Suppose the random sample I selected instead yields a patterning of correlations where the rank order of the sample correlations between the predictors and the outcome violates the rank order pattern of these three correlations in the population, like this:

	Y	X1	X2	X3
Y	1.00			
X1	0.35	1.00		
X2	0.42	0.52	1.00	
X3	0.33	0.48	0.44	1.00

Note that because of sampling error, the correlation between Y and X2 exceeds the correlation between Y and X1 but this is not the case in the population. The sample standardized regression equation now is

Y = 0.133 X1 + 0.290 X2+ 0.139 X3

and the narrative surrounding the magnitude of the predictor coefficients changes rather dramatically because of sampling error. It turns out that some forms of modeling are quite sensitive to such rank order reversals while others are not. Our goal as researchers should

be to minimize the amount of operative sampling error *throughout the entire covariance and correlation matrices* so as to preserve as best we can the rank ordering of correlation magnitudes. Otherwise, we can sometimes be seriously deceived.

The stability of correlation matrices in this regard is impacted not only by the sample size we use but also by the absolute magnitude of the correlations in the matrix because large absolute correlations have less sampling error than correlations near zero. Also, as the number of variables increases, the overall amount of sampling error in the matrix can accumulate. So, sample size, the magnitude of the correlations, and the sheer number of variables all can affect sampling error dynamics in complex ways. The sensitivity of the parameter estimates in your model to such sampling error also will depend on the nature and complexity of your model, further making it difficult to know how sample size impacts your modeling efforts.

You will find scattered in the literature a host of rules of thumb surrounding the sample size you need relative to the number of predictors, the number of estimated parameters, the number of degrees of freedom and/or the number of latent variables in your model. These recommendations are meant to address the above problem but they vary dramatically, such as recommendations of a ratio of 5 study participants per measured variable to as many as a ratio of 100 study participants per measured variable (Wolf, Harrington, Clark & Miller, 2013). Most Monte Carlo evaluations of the rules of thumb have found them to be wanting. A major limitation of them is that adequate sample size is not a simple function of the number of measured variables (MacCallum et al., 1999), so such ratios are, by definition, simplistic. Parenthetically, you also will encounter rules of thumb about sample size for the applicability of asymptotic theory, such as the need for samples sizes of at least 100 to 150 for asymptotic theory to hold. These rules of thumb also are gross oversimplifications.

IMPLICATIONS FOR SAMPLE SIZE DECISIONS

Many researchers think that the main driver of sample size decision making is statistical power. The choice is more complicated. Sample size can affect the amount of sampling error we have in our estimates. It also can affect the properties of estimators, such as their bias and efficiency. It can affect the applicability of asymptotic theory and the shape of sampling distributions, which affect accurate estimation of p values and confidence intervals. Effect size estimation is important for randomized trials, not just statistical significance. Sample size affects the magnitude of the margins of error surrounding effect size estimates. In SEM, sample size can affect the performance of global tests of fit, localized tests of fit, and tests of path coefficients for specific paths in a model, sometimes differentially so. Clearly, an analysis only of statistical power using assumption driven

canned software falls short when making informed sample size decisions. This is why I prefer, where possible, to use localized simulations to help choose a sample size. The present chapter will provide you with approaches for thinking about sample size selection taking all of the above factors into account.

In SEM, you can approach sample size decision making from the perspective of limited information SEM (LISEM) or full information SEM (FISEM). For LISEM, we often work on an equation by equation basis using an analytic method of our choosing (e.g., OLS regression, logistic regression, analysis of covariance). In these cases, standard software for power analysis often can be used to good effect, at least for gaining back of the envelope estimates of statistical power.

When I review grants for NIH that propose randomized trials, a common strategy that applicants use is to present a section on sample size in which they justify their chosen sample size in terms of the statistical power it yields for the effect of the treatment on the primary outcome. This is somewhat naïve not only because there is more than statistical power at stake but also because statistical power can differ for different statistical procedures. If the proposal has more than one analysis (which it usually does), than power analyses should be conducted for each analysis. A logistic regression with a binary outcome can be much more sample size demanding in terms of statistical power than OLS regression with a continuous outcome given comparable effect sizes. It, of course, is not practical to describe in a space constrained grant proposal power analyses for each equation or analysis. To deal with this, I often focus on the most sample size demanding analysis and do a thorough sample size analysis for it. If the sample size is adequate for it, it usually will be so for the other less demanding analyses.

Almost all power analyses we conduct are approximate and assumption driven. As you will see shortly, when conducting power or precision analyses in multivariate models, we often must make guesses about multiple parameter values in the population. These guesses are just that – guesses. Much of the software for power analysis assumes no missing data, that variables are normally distributed, and that the residual disturbances are homoscedastic. These assumptions often are violated in practice making the results of power analysis using canned software approximate at best. Analysts need to appreciate the approximate nature of power analyses. If practical constraints permit a larger sample size, I often use somewhat larger N than what software suggests.

SAMPLE SIZE AND STATISTICAL POWER

Suppose you are listening through a set of earphones and trying to decide whether you hear a particular signal. The static on the earphones makes this difficult for you. You have been

told that you should hear the signal within 30 seconds. One type of error you could make is to say you heard the signal when, in fact, it did not occur. This is analogous to a Type I error. Suppose making such an error would lead to negative consequences. You would want to be very sure of yourself. Only if you are virtually certain you heard the signal would you say you heard it. This is similar to setting a low alpha level in an investigation.

On the other hand, there is another type of error you could make–saying you did not hear the signal when, in fact, it was there. This corresponds to a Type II error. The ability not to miss the signal corresponds to the power of a statistical test. If you have a very sensitive ear, you will be likely to detect the signal when it occurs (high power); if you do not have a sensitive ear, you will be more likely to miss the signal (low power).

In this section, my concern is with power analysis for sample size decision making. I first describe two factors other than sample size that can affect statistical power because addressing them represent strategies for increasing statistical power in scenarios where increasing sample size is not viable. I then distinguish between conducting power analysis at the level of omnibus tests versus the level of more focused contrasts and underscore the importance of conducting power analysis at the level that is compatible with your statistical strategy. Finally, I walk you through the mechanics of using the canned power analysis programs on my website to assist you in gaining perspectives on statistical power.

Factors Affecting Statistical Power Other Than Sample Size

The Alpha Level

Using the electronics analogy from above, it is evident that the value of the alpha level affects the power of the statistical test. If you are very conservative about saying you heard the signal, then this decreases the likelihood you will say the signal is there when it is indeed present. The tradition of adopting a low or conservative alpha level in behavioral science research evolved from experimental settings where it was very important to avoid a certain kind of error. An example is testing a new drug with the aim of ensuring it is safe for the general adult population. In this case, deciding that a drug is safe when, in fact, it is not is an error that is certainly to be avoided. Under these circumstances, the proposal "the medicine is unsafe" would be cast as the null hypothesis with a low alpha level chosen so that the medical researcher has little risk of concluding the drug is safe (H_1) when actually it is not (H_0). By casting consequential errors as Type I errors in the framework of hypothesis testing and then setting a low alpha level, researchers minimize the risk of committing the error.

Many social scientists believe we have been preoccupied with Type I errors at the expense of Type II errors. The argument is that it is hard to justify that a Type I error will

have the drastic implications implied by a low alpha level relative to a perhaps more costly Type II error. Perhaps we need a better balancing of Type I and Type II errors instead of the common bias towards avoiding Type I errors even at the cost of insufficient Type II error control. My own bias is to avoid both types of errors and if I can set low probabilities of them occurring through my choice of sample size and covariate control, I do so. But I also am cognizant of the fact that that setting low probabilities of Type I errors can affect the probability of making a Type II error and that sometimes balancing the two is needed.

The Strength of the Effect in the Population Relative to Population Variability

Another factor influencing statistical power is the magnitude of the effect in the population. Statistical power is larger for strong effects than for weak effects, everything else being equal. Returning to our electronics analogy, it is much easier to detect a signal in a noisy environment if the signal is strong as compared to faint relative to the operative noise.

If we think of effect size and power in these terms, then we should be able to increase statistical power by eliminating noise from the system, thereby increasing the strength of the signal *relative to* the noise that is present. There are several strategies researchers can use to reduce noise. For example, gender impacts depression, with females showing higher levels of depression than males. Suppose we conduct a study to examine the effects of a new therapy on depression. We randomly assign people to either the new therapy condition or a control condition that does not receive the therapy. If the study includes both men and women, then in the presence of random assignment, both sexes will be represented in the treatment condition as well as the control condition. Within either condition considered separately, the presence of males and females will create variability in depression. This variability represents "noise" that detracts from the "signal" we are trying to detect, namely the effect of the treatment on depression. One way to eliminate this noise is to conduct the study only on women (or men). This would have the effect of removing biological sex as a source of noise in the study. The strategy, of course, has the disadvantage of reducing the external validity of the study. Alternatively, one can measure factors like biological sex and then statistically covary them out. This reduces the within-group standard deviations and, in turn, reduces noise. The statistical power of the test will be increased, accordingly.

The above point is important because it indicates strategies social scientists can use to increase power other than increasing sample size. These strategies become particularly important in research domains where practical constraints limit the number of research participants one can obtain. I elaborate this point below.

Omnibus Tests versus Focused Contrasts

When conducting a power analysis for methods such as analysis of variance and multiple

regression, many researchers conduct the power analysis on the omnibus or overall effect (e.g., the overall main effect for mean differences or the overall multiple correlation). However, few researchers stop at omnibus statistical tests when analyzing data; they pursue more focused tests. In ANOVA, researchers examine pairwise mean differences within a factor and in multiple regression, researchers examine individual regression coefficients associated with predictors. It is important to examine the statistical power associated with these more focused tests because they often will be the main source of conclusions in the study. Often, power for these contrasts is lower than that of the omnibus test.

The Mechanics of Power Analysis

In this section, I walk you through the software I provide on my website for traditional power analysis in the context of per equation LISEM analysis. These are my versions of "canned software power analysis" but are available to you at no cost. They share all the limitations of such programs more generally. They provide "back of the envelope" estimates of statistical power in a wide variety of scenarios.

To conduct power analysis, we need to specify the alpha level we intend to use (typically 0.05 with a two tailed test, which is the default in my software), the population effect size we want to target, and the minimum level of power we want to achieve (e.g., 0.90). With this information in hand, most software programs determine the sample size one needs to use to have sufficient sensitivity for the statistical test in question. The programs on my website allow you to create customized power tables with sample sizes of interest as rows, population effect sizes of interest as columns, and the table entries being power values. You can have as many rows and columns as you wish. I illustrate the usefulness of being able to construct customized tables shortly.

Specifying Target Population Effect Sizes

As noted, a basic facet of power analysis is specifying the strength of the effect in the population that the power analysis targets. For most program evaluation scenarios, the actual population effect size should **not** be what we specify when we conduct power analysis. Rather we specify what the minimum effect size value is that we want to be sure to detect as contrasted with effect size magnitudes that we don't care if we miss them. Stated another way, we are not obsessed with rejecting the null hypothesis per se; rather, we want to be sure that we can reject the null hypothesis *when doing so matters*.

Consider the case of salary discrepancies between males and females in academia. Suppose we want to test for discrepancies in the mean annual salaries as a function of biological sex and we want power of 0.90 to detect a difference. Suppose there is indeed a sex difference in annual income, but that the difference is $1. The null hypothesis of no

difference in annual salaries is not true, but do we really care if we miss such a trivial difference in our study of sex differences in annual salaries? Probably not. How about if the difference is $50? We still probably do not care if we miss this "effect" either. It is simply too small to worry about. But suppose the difference was $5,000? This is a considerable amount of money that will purchase quite a bit. We would not want to miss a discrepancy as large as this.

When specifying the population effect size to target, we want to specify the minimal effect size that we want to be sure to detect. Values below that effect size are too small to be of interest, but values equal to or larger than that effect size are important to detect. I refer to this as a **threshold value for effect size** (TES) or the **minimal important effect size** (MIES) and it is this value you want to use in your power analyses, not what you think the effect size in the population actually is. How do we know what the threshold value should be? There is no easy answer to this question, and the choice of a value must be dictated by substantive concerns and past research. As I discuss in Chapters 1 and 10, in my opinion, researchers give this matter too little attention. Instead, they often fall back on Cohen's (1988) classic criteria of what constitutes a "small," "medium," and "large" effect size despite the fact that Cohen's criteria are arbitrary and encouraged researchers not to use them (see Chapter 10).

Consider as an example a randomized trial where individuals are randomly assigned to two conditions, an active control group that receives educational materials about depression and a condition that applies a new form of cognitive-behavior therapy (CBT) to reduce depression. Suppose I want to compare the lost income due to days of missed work because of depression over a six month period for individuals in the CBT and the control conditions. Suppose the population average lost income difference between the two groups is $1. I would argue that this is trivial and I do not care if I "miss" this effect. On the other hand, if the true difference is $1,000, almost all would agree that this is meaningful. Should I use the value of $1,000 as my threshold? How about $500? What do you think the value should be and how can you justify it? Suppose I work with very low income, inner city populations who are economically disadvantaged. I might decide that for this particular population, a $100 disparity between the treatment and control groups over a six month period is important and therefore use it as my TES or MIES.

As I discussed in Chapter 10, I believe that for most areas of research, there will be a range of values either the scientific community or the organization that has hired you to evaluate their program would agree represents trivial effects, a range of values the most would agree represent meaningful effects, and a range of values that are in a "gray area" and which are debatable. I try to specify, as best I can, the values in these latitudes. In the present example, I might build a case for treating lost income between $60 to $180 (or

about $10 to $30 per month) as being in the "gray area." Saving less than $10 a month, I might argue, is not really going to matter all that much to the lives of people but saving more than $30 a month could be consequential. Thus, the $60 to $180 range is my "gray area." I might then decide to conduct the power analysis on the midpoint of the interval, namely $120 (or about $20 per month). Or, I might adopt a conservative approach and target an effect size that represents the smallest value of the latitude. A liberal approach is to use the largest value of the latitude. In the final analysis, I usually conduct power analyses for multiple values in the gray area and make sample size decisions, accordingly.

My point is that choosing a target effect size for power analysis is often a difficult, value-laden enterprise. I find it troublesome that some researchers treat effect sizes like t-shirts that come in sizes of "small," "medium," and "large," with the same definitions of these sizes applying to all situations. We need to be more nuanced than this.

Effect Size Sensitivity

A popular perspective on effect size in power analysis is to focus on a given sample size and then to specify the minimum population effect size that it will be reasonably sensitive to detecting. For example, suppose in the lost income example I decide that the lowest power that is acceptable to me is 0.80, which translates into a Type II error rate of 1-0.80 = 0.20. I decide to focus on a sample size of 125 per group because this is my "ceiling" sample size, i.e., it is the largest sample size I feel I can practically get given the resources available to me. Suppose I am interested in the effect size sensitivity of a sample size of 125 per group for a mean comparison of a continuous outcome between a treatment and control group when contrasted using an independent groups t test. I decide to isolate the value of the population effect size that will yield power of 0.80 for N = 125 for a two group t test of independent means. I might find that it equals $175. This is the **effect size sensitivity** of N = 125 per group for this test; for a sample size of 125 individuals per group, I will have reasonable sensitivity to detect an income difference of $175 or greater.

Some researchers, including myself, prefer to report power analyses using effect size sensitivities for different sample sizes when conveying the implications of using different N. For example, in grant proposals, I often provide to reviewers the sample size I have settled upon (or two or three alternative sample sizes I am considering) and then convey the effect size sensitivity of that sample size. I then build a case for why the sensitivity level is satisfactory.

Standardized or Unstandardized Effect Size Indices for Power Analysis

As discussed in Chapter 10, there are two types of effect size indices, unstandardized and standardized. Unstandardized effect size indices work in the raw metric of a variable, such

as the raw mean difference or an unstandardized regression coefficient. Standardized effect size indices impose a transformation on the unstandardized effect size to make it have a metric that some researchers find more meaningful, such as Cohen's *d* or a correlation coefficient. In general, most (but not all) of the utilities for power analysis I provide on my website work with unstandardized effect size indices, but you can adapt most of them to accommodate standardized effect sizes, as I show below.

Power Analysis for Mean Differences Between Independent Groups

In my programs for power analysis of mean differences between two or more independent groups (e.g., an intervention and control group), you specify the minimal size of the mean difference in the populations that you want to be sure to detect and what you think the population standard deviation is for each group after you control for covariates if covariates are part of your model. Your estimate of the population standard deviations typically will be informed by past research and/or a pilot study. Like other power analysis software, the traditional assumptions of normality and across group homogeneous variances are made. As a result, once you have specified the population standard deviation for one group, you have specified it for all groups. The power estimates also assume equal per group sample sizes, which is true of most power analysis software. If the per group sample sizes in your study are not or will not be equal, as long as they are not too unequal, the power estimates that assume equal group sizes will yield reasonable, ball-park power estimates. If your sample sizes will be quite discrepant, you can use the simulation strategies I discuss later to determine statistical power

In my programs, you can invoke Cohen's standardized criteria that define effect sizes as fractions of standard deviations by specifying the population within-group standard deviation as being 1.0 and then specifying the mean difference values as either 0.20, 0.50, or 0.80 for Cohen's small, medium, and large effect sizes, respectively. The resultant Ns will map onto those needed to achieve power for a raw mean difference equivalent to 2/10 of a standard deviation, half of a standard deviation, and 8/10 of a standard deviation, respectively, no matter what the actual value of the standard deviation is. If in your proposed study your outcome is measured in dollars and the population within group standard deviation is $10,000, the above strategy will tell you the sample size you need to detect a raw mean difference of $2,000, $5,000, and $8,000 with reasonable power.

The programs on my website create customized power tables for you. For the lost income example that contrasts the means of two groups, I use the program called *Power: Contrasts* because I am implicitly using formal contrast analysis. In the two-group case, this statistical test is the same as applying the traditional independent groups t test. The contrast approach specifies the contrast of interest using the following formula:

$$\psi = c_1 \mu_1 + c_2 \mu_2$$

where ψ is the contrast value of interest and to the right are the population means for each group multiplied by a contrast coefficient. The mean difference is defined using the contrast coefficients of 1 and -1:

$$\psi = (1) \mu_1 + (-1) \mu_2$$

which, when executed, subtracts the population mean for group 2 from the population mean for group 1. If I had 3 groups and wanted to contrast the mean for group 1 with the average of the means for groups 2 and 3, I would use the following contrast coefficients:

$$\psi = (1) \mu_1 + (-.5) \mu_2 + (-.5) \mu_3$$

The video and *details* link associated with the program describe the logic of the above specifications and provide multiple examples for defining contrast coefficients.

Here is a table from the program for the two group case where I have specified as input different population effect sizes as fractions of the within-group population standard deviation in the columns (because I set the population within group standard deviations to 1.0) and per group sample sizes of possible interest in the rows. The program then provides power estimates for each cell of the matrix, like this:

Custom table; rows are grp n, cols are mean diff, entries are power

	0.20	0.33	0.50	0.67	0.80
30	0.119	0.266	0.478	0.697	0.862
50	0.168	0.372	0.697	0.913	0.977
65	0.205	0.463	0.808	0.966	0.995
75	0.229	0.519	0.860	0.983	0.998
100	0.291	0.641	0.940	0.997	1.000
125	0.350	0.739	0.976	1.000	1.000
150	0.408	0.813	0.991	1.000	1.000

From this table, I see that for an effect size corresponding to a population mean difference equivalent to a third of a standard deviation (0.33), a sample size of about 150 per group yields statistical power of approximately 0.813.

If I want to use the table from the perspective of effect size sensitivity, I choose a group sample size of interest and then scan its row values to find the *a priori* specified power I want to impose, say 0.80. The value of the column effect size where the two intersect is the effect size sensitivity of the sample size. For example, for a per group sample size of 65 and setting my tolerance for a Type II error at 0.20 (or 1–0.20 = 0.80 power), the

effect size sensitivity of my test is $d = 0.50$ or greater, per the yellow highlighted portions of the table. Stated another way, given a group sample size of 65 per group, the independent groups *t* test will be reasonably sensitive (with power of 0.80) to population mean differences equal at least half a within-group standard deviation. Keep in mind that the table can be customized. Sometimes I regenerate my initial table with different row and/or column values to get a finer grained analysis of the dynamics at play as I zero in on the sample size I likely will use.

As discussed, one way of increasing power is to increase one's sample size. An alternative strategy if increasing N is challenging is to introduce covariates that are relatively uncorrelated with group membership (which will be true of any baseline covariate if random assignment to groups is used) but that impacts the outcome. If available, the baseline outcome or a proxy for it often is a good choice for such a covariate, per my discussion in Chapters 2 and 4. The program on my website *Power: Use of Covariates* helps you appreciate and convey to others the impact on power of using baseline covariates. For example, suppose without covariates the population mean difference between two groups equals a Cohen's d of 0.50 or half a standard deviation. The required sample size to obtain power of 0.80 is 64 per group when comparing the treatment and control groups. If I take into account covariates that account for, say, 40% of the variance in the outcome (which is a correlation of about 0.65 between the covariates and the outcome, which is not unheard of if the baseline outcome is the covariate), the program tells me that the required sample size to achieve power of 0.80 reduces to 38 per group.

On my website, I also provide a program for power analysis when comparing dependent or repeated measure means called *Power: One Sample*. The suite of programs, taken together, allow you to explore power analyses of tests of mean differences from multiple perspectives. Watch the video on my website for *Power: Contrasts* for multiple examples.

Power Analysis for a Regression/Path Coefficient

For multiple regression analyses, I provide a power analysis program for a regression coefficient in a multiple predictor context. You specify the effect size for a given predictor using a standardized index that, in my opinion, is more intuitive than that used by many power analyses programs. The index is the percent of unique explained variance associated with the predictor, such as the case where the target predictor accounts for 3% unique explained variance over and above the other predictors in the equation. This is specified in my programs by indicating what you think the population squared multiple correlation is when all of the predictors are included in the equation including the target predictor (e.g., 0.35) and then what happens to the squared multiple correlation when the target predictor

is dropped from the equation (the squared R reduces to 0.32). The change in R squared is an index of the minimum effect size for the predictor that you are interested in powering and represents the squared semi-part correlation associated with it.

As an example, I might conduct a power analysis to determine the required sample size to detect an effect for a predictor that accounts for at least 3% unique explained variance in a five predictor regression analysis. I set the population squared R in the population to, say, 0.25 when all 5 of the predictors are in the equation (or to any other value I think maps onto my analytic scenario). To examine power for a predictor that accounts for 3% unique explained variance, the squared R would drop to 0.22 if I eliminated the predictor. For this scenario and an alpha level of 0.05 (two tailed test), the required total sample size to achieve power of 0.80 is about 200.

It turns out the power value is influenced by the absolute magnitude of the two squared multiple correlations, so if I enter squared Rs of 0.40 and 0.37, I will get different results than if I enter 0.25 and 0.22. You need to make reasonable guesses of the magnitudes of the two population squared correlations. Power also will be affected by the number of predictors in the equation, with more predictors producing less power, everything else being equal, though the effect usually is trivial when N is large.

Here is an example of a table I generated exploring different effect sizes for a 7 predictor equation in which the population squared R for the full equation is 0.30. The columns were 0.27, 0.25, 0.20 and 0.15. These entries represent 3% unique explained variance, 5% unique explained variance, 10% unique explained variance, and 15% unique explained variance, respectively, because the values reflect how much the squared R might decrease when the target predictor is dropped. The sample sizes in the rows are N = 100, 120, 140, 150 and 160. The table is:

R square for full equation: 0.3

	0.27	0.25	0.20	0.15
100	0.510	0.727	0.952	0.993
120	0.592	0.808	0.979	0.998
140	0.662	0.867	0.991	1.000
150	0.694	0.890	0.995	1.000
160	0.723	0.909	0.997	1.000

For a predictor that has unique explained variance corresponding to 10%, a sample size of 100 will have power of 0.952 for the test of the target coefficient.

From an effect size sensitivity perspective, suppose I decide my sample size ceiling is 120 and that I can't practically get a larger N than this. I want to have my Type II error rate be 0.20 which means I want power of 0.80. I look at entries in the row for the sample

size of 120 and find the entry that equals 0.80. I highlight it in yellow. The effect size sensitivity for a sample size of 120 is unique explained variance of 5% or greater (which is the difference 0.30 minus 0.25 multiplied by 100. The video associated with the regression coefficient power program on my website (called *Power: Regression coeff*) walks you through several examples.

You can also conduct an approximate power analysis for a squared correlation coefficient using the *Power: Regression coeff* program. You specify 1 predictor and provide the squared correlation for that predictor, say 0.05. If you eliminate the predictor, the squared correlation becomes 0. Here is the table that results for various sample sizes for a squared correlation of 0.05:

```
R square for full equation: 0.05

         0.00
100     0.622
120     0.703
140     0.769
150     0.797
160     0.822
```

An N of 150 will have power of approximately 0.797 for a test of a population squared correlation of 0.05.

Power Analysis for a Logistic Coefficient

Software programs for power analysis of logistic regression coefficients, including the programs on my website, treat the case of a continuous predictor as distinct from the case of a binary predictor. For my program to determine statistical power for a continuous predictor (called *Power: Logistic coeff 1*), the population effect size for the predictor, X, is stated as an odds ratio. This is the multiplicative factor by which the odds of Y change given a one unit change in X, holding constant all other predictors. As discussed in Chapter 5, the odds ratio is the exponent of the population logistic coefficient associated with the predictor.

For the program on my website, the target predictor defaults to an X metric consisting of a mean of 0 and a standard deviation of 1.0, i.e., it mimics a standardized metric. You can change this value if you want but using a standardized metric gives the analysis broader applicability. The population multiplying factor (or odds ratio) you provide is for a one unit change in X which corresponds to a one SD change in X given the use of the standardized metric. For example, an odds ratio value of 1.5 means that every time X increases by one standard deviation, the odds of engaging in the event changes by a multiplicative factor of 1.5. If the odds of engaging in Y is 2.0, then when X increases by one unit (standard

deviation), the odds of Y changes by a multiplicative factor of 2.0. And so on.

Suppose I have 5 predictors of a binary Y. In the program on my website, I need to specify (a) the sample sizes I am considering, (b) the odds ratios of interest to reflect the effect size, (c) the population squared correlation between my target predictor and the other four predictors in the model (to index multicollinearity), (d) what I think the population *event rate* is for Y when all predictors equal their means (this is input as the probability of Y or the proportion of people who engage in Y when all predictors equal their mean), and (e) the alpha level. Suppose the squared correlation for collinearity is 0.15, the event rate at the predictor means is 0.50, and the alpha level is 0.05. I might explore population odds ratios of 1.2, 1.5, 1.8 and 2.0 for sample sizes of 150, 175, 200, 225, 250, 275, and 300. Here is the table that results:

Custom Table: Row is total N, column is odds ratio, entries are power values

	1.20	1.50	1.80	2.00
150	0.176	0.629	0.913	0.975
175	0.198	0.696	0.948	0.988
200	0.220	0.753	0.969	0.995
225	0.242	0.801	0.982	0.998
250	0.264	0.840	0.990	0.999
275	0.286	0.873	0.994	1.000
300	0.307	0.899	0.997	1.000

For a coefficient that reflects an odds ratio of 1.80, a total sample size of 150 will have statistical power of 0.913. From an effect size sensitivity perspective, suppose I have a sample size ceiling of 225. Using the above table and for a Type II error rate of 0.20 (power = 0.80), the effect size sensitivity of a sample size of 225 is an odds ratio of 1.50 or greater.

The program *Power: Logistic coeff 2* does a power analysis similar to the above but for a binary predictor. In this program, the effect sizes are specified using proportions rather than odds ratios. As an example, I might specify the population control group proportion of those who perform Y to be 0.50. I specify for the columns of the table the possible population proportions for the intervention group of 0.55, 0.60, 0.638, 0.64, 0.65, 0.70, 0.75. Each of these proportions define a different effect size in terms of a proportion difference that subtracts the control group proportion (in this case, 0.50) from the intervention group proportion. For example, the value 0.55 represents a 0.05 proportion difference of 0.55-0.50; the value 0.60 represents a 0.10 proportion difference of 0.60-0.50; and so on. Note also that if you want, you can translate these proportion differences into odds ratios. For example, for the proportions 0.55 versus 0.50, the odds ratio is [0.55/(1-0.55)] / [(0.50/1 – 0.50)] = 1.22. I set the squared correlation for collinearity to be 0.0 because the intervention versus control group assignment is random which means the

binary predictor should be uncorrelated with the other predictors. I explore sample sizes of 200, 250, 300, 350, and 400 per group and an alpha level of 0.05. Here is the table that results from the program:

	0.55	0.60	0.638	0.64	0.65	0.70	0.75
200	0.170	0.521	0.799	0.810	0.861	0.984	0.999
250	0.201	0.615	0.878	0.887	0.926	0.996	1.000
300	0.232	0.694	0.929	0.935	0.962	0.999	1.000
350	0.263	0.759	0.959	0.964	0.981	1.000	1.000
400	0.294	0.813	0.977	0.980	0.991	1.000	1.000

To detect a 0.10 difference in proportions (0.60-0.50) with power of approximately 0.80, I need about 400 individuals per group. To detect a proportion difference of 0.14 with power near 0.80, I need a sample size of approximately 200 per group. Keep in mind that by using control and intervention population proportions near 0.50, sample size demands are greater than if the population proportions were towards their extremes, 0 or 1.0, vis-a-vis the principles I discussed on the effect of variability on sampling error.

From an effect size sensitivity perspective, suppose I have a sample size ceiling of 200 per group. Using power of 0.80, a two tailed test, and multicollinearity of 0, I find from the above table that the effect size sensitivity is 0.638 – 0.500 = 0.138 or greater (see the yellow highlighted table entries).

With practice and by having the ability to create your own custom power tables, you can conduct reasonable power and effect size sensitivity analyses for a wide range of research scenarios. Watch video C on my website for illustrations.

Power Analysis for Selected SEM Tests

Power Analysis for the Global Chi Square Test. A commonly used test in full information SEM is the chi test of global fit (see Chapter 7). The test evaluates a null hypothesis of perfect model fit in the population (or a zero residual matrix) against an alternative hypothesis of non-perfect model fit (or a residual matrix with at least one non-zero element). To the extent that you rely on this statistic, you will want to ensure you have sufficient statistical power or test sensitivity to detect meaningful ill fit in the population model. I provide a program on my website, called *Power: SEM chi square test*, to conduct such power analyses to help you choose a sample size that will provide sufficient power/sensitivity for the global chi square test of fit.

Like other power analyses, you need to specify a minimum population effect size that you want to be sure to detect. This takes the form of an *a priori* specified degree of discrepancy from perfect model fit. The tradition for power analysis is to express this disparity in RMSEA units. Recall that the RMSEA has a lower bound of zero (which

indicates perfect model fit) and that larger values indicate a worse fitting model. Usually, RMSEA values fall between 0 and 1 and they tend to be lower than 0.25. In the SEM literature, RMSEA values less than 0.08 are often said to reflect satisfactory model fit and values less than 0.05 are deemed as good model fit. However, there is controversy about these standards and many methodologists argue (persuasively) against their use. The problem with relying on the RMSEA to specify/quantify population ill fit for purposes of power analysis is that it uses a counterintuitive metric for a statistic that is complexly determined. My preference instead is to conduct global chi square power analyses using the localized simulation strategy that I describe later, but for those who gravitate towards the RMSEA, I provide a program for a power analysis of the global chi square test.

In the program, suppose I set the minimum degree of population model misspecification that I want to be sure to detect to be an RMSEA population value of 0.08 for the misspecified model. For the program, I need to specify the model degrees of freedom. You can derive this algebraically or you can run your model using Mplus with hypothetical data of any sample size and then check the Mplus output for the model degrees of freedom reported just under the chi square test on the output. For my program, you also specify the power you desire (e.g., 0.80). Given this information, the program will calculate the sample size you need to achieve that power for the traditional chi square test as well as power values a little lower and a little higher than your input power value. For example, if my model has 15 degrees of freedom and I desire power of 0.80 to detect a population RMSEA of 0.08 or greater, here is the program output:

```
                    Results
n for power = 0.8       197
n for power = 0.75      179
n for power = 0.80      197
n for power = 0.85      219
n for power = 0.90      247
n for power = 0.95      291
```

The first entry is the required sample size for the power value you requested; I need a sample size of 197 cases to have adequate power to detect model misfit of RMSEA = 0.08 or greater. The remaining output is self-explanatory. Note that this program is only applicable to single group models using maximum likelihood estimation. If you have a multigroup model or use a robust estimator, use the simulation approach discussed below.

Power Analysis for the Chi Square Difference Test. I also provide on my website a program called *Power: Compare models* to conduct a power analysis for a chi square difference test for nested models using maximum likelihood estimation. I refer to one of the models as the constrained model and the other as the unconstrained model. The null

hypothesis posits identical fit for the two models in the population. The alternative hypothesis is that the unconstrained model fits better than the constrained model (the test itself is described in Chapter 8). You specify in RMSEA units the model fit difference between the constrained and the unconstrained models that you want to power against. Instead of specifying a single RMSEA difference, such as a difference of 0.05, you must specify the component parts of that difference, namely the presumed population RMSEA for the constrained model (e.g., RMSEA = 0.10) and the presumed population RMSEA for the unconstrained model (e.g., RMSEA = 0.02). You also specify the degrees of freedom for each model. The constrained model always will have the larger RMSEA and the larger degrees of freedom. Like the power analysis for a global chi square test of fit, the challenge is that the method uses RMSEA units which are counterintuitive.

As an example, the constrained model presumed population RMSEA might be 0.10 with 12 degrees of freedom and the unconstrained model population RMSEA might be presumed to be 0.02 with 10 degrees of freedom. Here are the results from my program for a desired power of 0.80:

```
                   Results
RMSEA difference      0.08
n for power = 0.8    84.00
n for power = 0.75   67.00
n for power = 0.80   84.00
n for power = 0.85   95.00
n for power = 0.90  110.00
n for power = 0.95  134.00
```

I need a sample size of about 84 cases to have adequate power (0.80) to detect the model difference using the chi square difference test. As with the single model chi square test, this program is only applicable to single group models with maximum likelihood estimation. If you have a multigroup model or you want to evaluate power using a robust estimator, use the localized simulation approach I describe below.

Additional Power Analysis Programs. My website has several other power analysis programs that may be of use to you, including power analysis for tests of close fit and for confirmatory factor analyses.

The Role of Pilot Studies and Past Research in Power Analysis

Some methodologists believe it useful to inform power analysis by using documented effect sizes from prior research or pilot research. Although there are scenarios where this is appropriate, it often is not. A researcher might conduct a meta-analysis of past research to identify the average effect size for the intervention of interest. This average effect size

is then used to power the proposed study. To me, what matters most is specifying a minimum effect size magnitude that you want to be sure to detect. Whether that effect size has been observed in prior research is not necessarily germane. If I decide detecting a $750 disparity in annual salaries between males and females is important but prior studies with similar populations have observed, on average, a $1,500 sex difference, I am not going to power my study relative to an effect size of $1,500 if I have decided a difference of $750 is meaningful. There are, of course, cases where effect size choices are appropriately informed by prior research. Suppose that current medications for an illness lead to a 60% cure rate. I want to introduce a new medication that is less expensive and has fewer side effects. I might then power my study to detect cure rates that are at least near 60% so I can take advantage of the reduced costs and fewer side effects of it.

Pilot research also is of questionable value if it uses a small sample size to define the likely or desired effect size. A small N in pilot research makes the effect size estimate from that research subject to large amounts of sampling error, often resulting in untrustworthy estimates (Kraemer et al., 2006). For example, the margin of error for a Cohen's *d* statistic based on a pilot N of 20 is about ±0.90; for a sample size of 40 it is ±0.63; and for a sample size of 60 it is ±0.52. Such estimates are not very reliable and one would be hard pressed to justify them as a basis for powering a major, expensive clinical trial.

Pilot studies can be used to estimate or inform choices about the population standard deviation, base rates for proportions, predictor collinearity, and distribution normality (Leon, Davis & Kraemer, 2011) that need to be taken into account when making sample size decisions. However, relying on pilot studies to do so also can suffer from the small N problem and sampling error for these goals as well.

When a researcher relies exclusively on an effect size from a pilot study or an average from prior research to conduct a power analysis, the rationale usually is based on an antiquated view that one's primary goal in power analysis is to reject the null hypothesis even if the true effect size is trivial and without implications. More modern views have shifted perspectives towards not just powering a study to detect *any* effect no matter how small to one of detecting meaningful effect sizes for the research questions being asked. Making such meaningfulness judgments usually requires far more than importing an average effect size from a meta-analysis or from estimating an effect size in a small N pilot study.

Post Hoc Power Analysis

Some statistical software reports post hoc power analysis that calculates the power of a contrast or coefficient based on the effect size observed in the data being analyzed. This often is referred to as **post hoc power analysis**. The approach is problematic because (1)

it assumes the minimal meaningful effect size equals the effect size observed in the study, and (2) it fails to appreciate that the observed effect size is subject to sampling error. Having said that, it is not unreasonable to conduct certain kinds of post hoc power analyses to evaluate the operative power in a study, especially if the authors of the study are making important conclusions based on statistically non-significant effects. In such power analyses, one should not use the observed effect size in the study but rather a logically derived minimum important effect size to set the standard for power analysis. I discuss other applications of post hoc analyses to assist in data analysis below.

Power Analysis for Robust Statistics

Wilcox (2021) describes robust analytic methods (e.g., MM regression, quantile regression, trimmed mean analysis) that can be used in LISEM in place of more traditional methods when evaluating RET models. These methods usually do not have power analysis software associated with them. A crude approach for power analysis for these methods is to use power analysis software for traditional non-robust methods with the idea that the results generally (but not always) will be indicative of power for their robust counterparts, indeed conservatively so. Alternatively, one can conduct localized simulations for the robust methods using R, but this requires knowledge of R programming.

Power Analysis for Group Administered Interventions

Sometimes we conduct randomized trials where instead of assigning individuals to conditions, groups of individuals (often called **clusters**) are assigned to conditions and the groups receive a common experience or intervention. For example, a treatment protocol for child anxiety might involve forming small groups of children with each group then receiving an intervention of group activities designed to reduce anxiety. A control condition might assign children to groups, but the groups engage in activities that have nothing to do with anxiety reduction. The use of groups or "clusters" introduces analytic complications per my discussion in Chapter 25. When conducting power analysis, these cluster or group effects need to be accounted for. I do not provide canned software on my website for conducting clustered power analyses although I offer a program called *Power: Cluster adjustment* that can give you a rough back-of-the-envelope appreciation of how clustering affects statistical power in some contexts. The best way to gain perspectives for statistical power in such contexts is to do localized simulations in Mplus, which I describe below. For the program on my website, I use the approach by Donner et al., (1981).

Concluding Comments on Sample Size and Statistical Power

In sum, sample size affects our ability to detect meaningful effects in RETs. We can specify

a priori how sensitive we want our tests to be and then choose sample sizes that achieve that sensitivity. Canned power analysis software can help us select sample sizes to achieve our goals but these programs are only approximate and they are incomplete. They make a host of assumptions, some of which are unrealistic in practice. An alternative is to choose sample size based on localized simulations, a strategy I describe below. If the sample size demands of your study are unworkable because of pragmatics, there are other strategies you can use to increase statistical power. Probably the most viable strategy is to use thoughtfully selected covariates to reduce "noise in the system." With carefully chosen covariates, you might be able to cut your sample size in half, without sacrificing power.

An important task for power analysis is to specify the minimum effect size that you want to be sure to detect in your study. Some effects are trivial and it does not matter if you miss them. Others are substantively meaningful and you want to be certain you do not miss them. The effect size sensitivity of the sample size you settle upon should be reasonably capable of detecting minimum meaningful effects.

The default for many researchers is to seek power of 0.80, which means you have a 20% chance of missing a meaningful effect. The default also is to use an alpha level of 0.05, which means you are willing to tolerate only a 5% chance of concluding an effect is present when, in fact, it is not. There is a notable asymmetry in error tolerance for the two types of errors in hypothesis testing (Type I and Type II errors). When choosing sample sizes, you need to justify in your mind such asymmetry and think through the consequences of both types of errors. Set your target error rates accordingly.

Power analyses should be pursued in the context of the broader RET model you are using. There usually will be multiple equations in your model. If you use LISEM or piecewise SEM, separate power analyses need are applied to each equation. In an RET, this takes the form of power analysis for the effects of the treatment on the mediators, for the effects of the mediators on the outcome, and for the independent effects of the treatment on the outcome over and above the mediators. Your approach to power analysis can differ depending on whether you use LISEM or FISEM, but often the sample size needs converge in the two methods. However, there are definite exceptions, some of which I discuss below.

SAMPLE SIZE AND MARGINS OF ERROR

Although most social scientists perform power analyses before they conduct studies to ensure that their sample sizes are adequate, sample size decisions also can be approached from the perspective of margins of errors (MOEs), i.e., obtaining small enough MOEs about one's estimates. For example, if I tell you that I plan to estimate a correlation between two variables and that based on my sample size the MOE for the correlation estimate will

be ±0.60 correlation units, you would likely be critical of my study plans because the MOE will be too large. On the other hand, if I tell you that my sample size is such that the MOE will be plus or minus 0.01 correlation units, then you likely will be satisfied that the estimate will be informative. Many factors impact the magnitude of a margin of error, but a major influence is sample size; the larger the sample size, the smaller the margin of error, everything else being equal.

When planning an experiment, we can think about how imprecise we are willing to let our estimate be, i.e., how large a margin of error we are willing to tolerate. On my website, I provide programs to determine the sample size necessary to achieve a MOE no greater than a given size for various statistical methods.

The value of the margin of error observed in data for a given study will vary from one sample to the next because of sampling error. This means that you cannot be certain that when you conduct your study, you will get the exact MOE value you planned based on the sample size provided by my programs. For example, if you seek to estimate the mean annual income in a population and you want a margin of error no greater than ±$1,000, the programs provide you a sample size to attain that margin of error. However, when you collect your data and calculate your MOE using that sample size, the margin of error may be somewhat larger (or smaller) than $1,000. The concept of an **assurance probability** (also sometimes called **tolerance**) refers to the probability that the MOE will, in fact, be no larger than the MOE you specified in the sample size determination algorithms. When using programs to determine sample sizes for MOEs, you need to choose an assurance probability with respect to the MOE.

Let's consider an example to make this concrete. Suppose I have a diet/exercise program designed to help people lose weight. I fit a linear equation that predicts weight in pounds from the treatment condition (1 = treatment, 0 = control) and a set of covariates, one of which includes the baseline weight of the individual. When I compare the treatment and control conditions, the regression coefficient for the treatment dummy variable reflects the covariate adjusted mean weight difference between the treatment and control groups. There will be a margin of error associated with this estimate. How large do I care that the margin of error is? Suppose I decide I want the MOE to be no larger than plus or minus 2 pounds. I use the program on my website called *MOE for regression* and enter the required information to calculate the sample size I need to use in my study to achieve a MOE of ±2.0 for the posttest weight difference. This includes providing information or guesses about the population standard deviation of the target predictor (which in this case is the standard deviation of the dummy variable for the treatment condition; for a two group dummy variable with dummy coding, it is 0.50), the population standard deviation of the outcome (suppose it is 15 pounds), the presumed squared correlation of the target predictor

with the other predictors in the prediction equation to index multicollinearity (which will equal zero if the target variable is a randomly assigned treatment condition variable), what I think the overall population squared correlation is between the outcome and all of the predictors in my regression equation (e.g., I might set it to 0.55, which is large because I include baseline weight as a covariate in my model), the number of predictors in my regression equation (there were 4), the percent of the confidence interval I want to use to define the MOE (95%), and the assurance probability, which I might set to 0.80. Here are the results from my program:

	MOE	Required N
Case 1	1.0	1616
Case 2	1.5	736
Case 3	2.0	425
Case 4	3.0	202
Case 5	4.0	122

To obtain a MOE of ±2.0, I need a total sample size of about 425, or 212 per group. The program also provides sample sizes for MOEs close to the one I requested to provide additional perspectives. For example, if I am willing to tolerate a MOE of ±3.0 pounds, I can use about 100 people per group instead.

How large should your MOE be? There is no simple answer to this question and it depends on how much error you are willing to tolerate given the substantive questions being asked. When working with percents, it is not uncommon to see national polls with MOEs between 3% and 5%. Polls whose MOEs are above 8% are seen as dubious. Some researchers import Cohen's guidelines for small and medium effect sizes to the specification of MOEs whereby you do not want a MOE larger than, say, 0.20 *d* units. But such an approach is arbitrary. I recommend you think carefully about tolerable MOEs given the consequences of large MOEs based on what you are studying.

I find it helpful to think in terms of MOEs rather than solely statistical power when making decisions about sample sizes. The perspective is different. In power analysis, we seek to specify an N that will allow us to find a statistically significant effect if a meaningful effect is present in the population. In MOE analysis, we approach the matter from the perspective of how much error we are willing to tolerate in our estimates. I describe below how to gain perspectives on both power and margins of error using localized simulations.

Some Technical Matters

In this section, I point you to statistical literature for determining sample sizes for desired levels of MOEs. You can skip this section if you are uninterested in such details.

Sample size and confidence interval estimates for mean comparisons use classic single degree of freedom contrast strategies in conjunction with precision analyses. These methods are discussed in Kupper and Hafner (1989) and Pan and Kupper (1999).

When estimating sample sizes for MOEs for correlations, the required sample size differs depending on the magnitude of the correlation in the population; in general, larger correlations are less sample size demanding. Kupper and Hafner (1989) suggest using a conservative approach for sample size determination in which you base your decision about sample size by always focusing on an expected population correlation of zero. This will then conservatively yield the largest required sample size for your margin of error relative to other expected correlations.

For regression coefficients, the sample size needed to obtain an interval width of a specified size has been described in Kelley and Maxwell (2003).

For percentages and percentage differences, one can use a variant of the approaches described in Kupper and Hafner (1989) and Pan and Kupper (1999). Like correlations, the required sample size differs depending on the magnitude of the percentage in the population; in general, the closer percentages are to 50, the more sample size demanding they are. Kupper and Hafner (1989) recommend estimating required sample sizes assuming the population percentage is 50, where sampling error is greatest. This will then yield a conservative estimate.

Kelley and Maxwell (2003, 2008) present approaches for estimating the width of confidence intervals for multiple correlations and regression coefficients (see also the work of Jiroutek, Muller, Kupper and Stewart, 2003). Satten and Kupper (1990) describe methods for sample size estimation for margins of errors for odds ratios (i.e., for the case of dichotomous outcomes; see also the Erratum by Kupper, 1990). Cesana, Reina, and Marubini (2001), Samuels and Lu (1992), and Bromaghin (1993) discuss such approaches for percentages and/or percentage differences. Algina and colleagues (Algina and Olejnik, 2003; Algina, Moulder, and Moser, 2002) apply sample size planning for margin of errors for correlations, partial correlations, the difference between squared multiple correlations, and squared semi-part correlation analysis. A general review of relevant literatures on sample size planning and margins-of-error is in Maxwell, Kelley, and Rausch (2008).

LOCALIZED SIMULATIONS FOR SAMPLE SIZE DECISIONS

When planning an RET, it can be helpful to conduct a localized Monte Carlo simulation to inform choices about sample size. I refer to this activity as conducting a **localized simulation study** because the simulation provides feedback about your specific study conducted under specific conditions. The material I cover here is a bit advanced but the

skills are well worth mastering. I begin by first describing the core logic of a simulation study using a comparison of two group means as the target method of analysis, a method you typically associate with an independent groups t test. I then expand on that core logic to describe how to analyze power for a wide range of statistical tests relevant to RETs using simulations. I augment this presentation with even more examples on my website.

Suppose I want to evaluate the effects of violating normality assumptions on statistical power when comparing group means. To do so in a simulation, I create two very large populations of normally distributed scores on my computer in a way that I know that the standard deviation of the Y scores in each population is 1.0, with the mean in population A being 0.50 and in population B it is 0.0. Thus, the mean difference is 0.50 and given that the standard deviations both equal 1.0, this maps onto a Cohen's d of 0.50. Using a traditional power analysis program, I find that the power for rejecting the null hypothesis using an independent groups *t* test is 0.80 when my sample size equals 65 per group. However, suppose I also generate a set of such scores on my computer, but I add a twist. In both groups, the population scores are non-normally distributed with skewness = 2.5 and kurtosis = 4.0. What happens to the power of the test of mean differences in this case? It turns out that Mplus has random number generators that allow me to generate such populations with characteristics that I desire so that I can do a deeper analysis to see what happens.

Suppose after creating the populations, I select a random sample of 65 individuals from each of them. I use Mplus to calculate a test of the difference between the two means based on the robust maximum likelihood algorithm that uses Huber-White estimation. This is not the same as an independent groups *t* test because Huber-White estimation on Mplus is based in asymptotic theory and it has robust properties. This fact causes me to wonder what the statistical power will be for this analytic scenario. I note whether I reject the null hypothesis for the test I conduct on the 130 individuals I have randomly selected. I fully expect to obtain a statistically significant result because, after all, the null hypothesis is false; the population means are not equal. Suppose I repeat this sampling process 20,000 times followed by a test of the mean differences each time. I note for each case if the null hypothesis is rejected. These repeated tests are called **simulation replicates** or, more simply, replicates. In executing this step, I essentially conduct 20,000 study replications but each study uses a different random sample of 65 cases per group and all of the tests are automatically performed by Mplus on my computer. I then calculate the proportion of times across the 20,000 replicates that the null hypothesis was rejected. This represents the power of the test. If the test is robust to violations to non-normality and the Huber-White estimator works well, I should find that I reject the null hypothesis about 80% of the time per what I found with the power analysis software. I can also calculate a confidence interval in each

replicate and count the proportion of times that the population group difference was contained in the interval across the 20,000 replications. This should be 95% of the time for a 95% confidence interval. The proportion of times the population value is in the interval is called the **confidence interval coverage**.

If my statistical power in this scenario was, in fact, near 0.80, then I would have increased confidence that I can use a sample size of 65 per group to compare means using the SEM based strategy implemented by Mplus even in the face of the non-normality that was present in the respective populations. The simulation study gives me perspectives that canned power analysis software cannot because the latter only provides power estimates for the case of homogenous variances with normally distributed data for a *t* test.

To teach you Mplus programming for simulations, I use here a simplistic example in which I have a treatment versus control condition, a single mediator, and a single outcome. I address more complex models later. I seek to conduct a pre-study power analysis for the model in Figure 28.1. The intervention teaches students study skills for math. There are two treatment conditions, an intervention and control group. The mediator is a test of student study skills measured four weeks post- intervention scored from 0 to 100 with higher scores indicating better skills. Students typically score about 60 on the test, with a standard deviation of 15 or so. The outcome is performance on the final math exam. Scores on the exam range from 0 to 100 with higher scores indicating better performance. Like many school exams, a score of 90 is excellent, a score of 80 is above average, a score of 70 is average, and so on. The standard deviation on the exam is usually about 15.

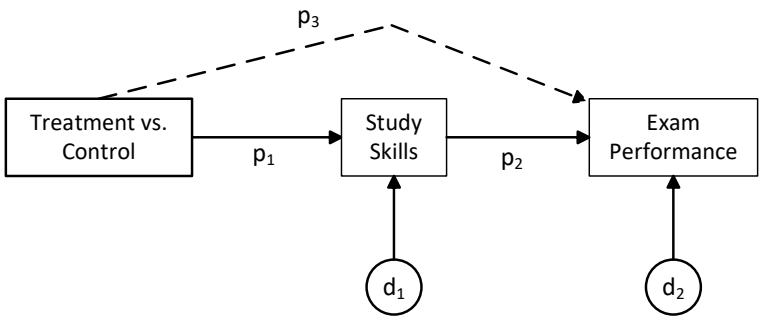


FIGURE 28.1. Simulation Example with Single Mediator

The model is unrealistic in its simplicity but serves my purposes for exposition. It is overidentified given it omits path p_3 (the dashed arrow); the effects of the program on exam performance are assumed to be fully mediated by program effects on study skills. This is

not an unreasonable assumption in practice. I omit baseline measures but introduce them later. My tentative plan for the study might be to use a sample size of 75 per group in the treatment and control conditions and I set up my simulation accordingly. I explore the properties of other sample sizes later.

Choosing Parameter Values

The first step in a simulation is to specify values of the population parameters that are needed to conduct the power analysis. In Figure 28.1, I seek to evaluate statistical power for paths p_1 and p_2. To do the analysis, I need to specify the population path coefficient values for p_1 and p_2 that I want to target or be sensitive to with adequate power, i.e., the minimal important effect sizes for each of them. It turns out that I also will need to specify the variances or standard deviations of each variable in the model, and values of the two disturbance variances because all of these also can affect my power estimates. As you will see shortly, I do not need to specify population means or intercepts because they are not used in this particular model. However, there are models where such specification is necessary. In the current case, I want to choose values for the path coefficients that reflect the minimal meaningful effect sizes that I want to be sure not to miss and then determine the sample size I need to obtain the desired power. I will work with the traditional alpha level of 0.05 for a two tailed test.

Simulation designers specify population parameter values using either the raw metrics that characterize model variables (e.g., study skills and exam performance each vary from 0 to 100) or they use a metric that is convenient, easy to work with, and that will produce the same core results as raw metric specifications. A common strategy for the latter approach is to treat all continuous measures as having an overall mean of zero and a variance of 1.0, much like a standardized metric. To be sure, the metrics are not formally standardized but by having a mean of 0 and a standard deviation of 1, the parameters behave much like standardized parameters that many researchers are more comfortable with. I use the raw metric approach here for simulation design but I show you the standardized metric approach in the Appendix. I prefer the former but you may find situations where the latter is easier to work with.

When specifying parameter values, it is important that they be logically consistent with each other. There are several formulae and regularities that will be useful in this regard. First, in the models we typically work with, the population variance of an outcome can be partitioned into two components, (a) systematic or between group variance and (b) error or within group variance. Here are some expressions of this partitioning that you may have encountered in your statistics courses:

$$var_{TOTAL} = var_{BETWEEN} + var_{WITHIN} \quad [28.1]$$

$$var_{TOTAL} = var_{REGRESSION} + var_{ERROR} \quad [28.2]$$

$$var_{TOTAL} = var_{SYSTEMATIC} + var_{ERROR} \quad [28.3]$$

$$var_{TOTAL} = var_{EXPLAINED} + var_{UNEXPLAINED} \quad [28.4]$$

A population parameter called eta squared or R squared makes use of the above partitions to define the proportion of the total variance in the outcome that can be accounted for by the predictors in a linear equation that predict the outcome, like this:

$$Eta^2 = var_{BETWEEN} / var_{TOTAL} = var_{REGRESSION} / var_{TOTAL} =$$

$$var_{SYSTEMATIC} / var_{TOTAL} = var_{EXPLAINED} / var_{TOTAL} \quad [28.5]$$

I will make use of this formula. Another useful formula for simulation design is a different type of variance partitioning but focused on a linear equation. Consider a linear model with a single predictor:

$$Y = \alpha + \beta X + \varepsilon$$

It can be shown through algebraic manipulation that

$$var(Y) = \beta^2 var(X) + var(\varepsilon) \quad [28.6]$$

or that the variance of the outcome Y equals the value of the squared regression/path coefficient for the predictor times the variance of the predictor plus the error (or disturbance) variance. As you will see, I also make use of this formulation when designing a simulation.

A final set of formulae I sometimes use focuses on the relationship between correlations and covariances. A correlation between two variables, X and Y, can be expressed as a function of a covariance as follows:

$$r_{XY} = (cov_{XY}) / [(SD_X)(SD_Y)] \quad [28.7]$$

where r_{XY} is the correlation between X and Y, cov_{XY} is the covariance between X and Y, SD_X is the standard deviation of X, and SD_Y is the standard deviation of Y. Using simple algebraic manipulation of Equation 28.7, a covariance can be expressed as a function of a correlation as follows:

$$cov_{XY} = (r_{XY})(SD_X)(SD_Y) \quad [28.8]$$

With these formulae in mind, let's work through the parameters I need to specify for the model in Figure 28.1. In the current example, I will use some but not all of the formulae.

First, working from left to right in the figure, I need to specify the variance of the dummy coded treatment condition, T, where 1 = the intervention group and 0 = the control group. If I have an equal number of participants in each group per random assignment, the population standard deviation of T always will be 0.50 (which is the standard deviation of a binary dummy variable with equal n per group) and the variance will be the square of this value, 0.25.

Based on my experience with both the study skills and exam performance measures, (which I hereafter call M and Y, respectively, for "mediator" and "outcome"), I decide to set each of their standard deviations to 15 which, as I noted, is commonly found in the literature. The variances are the square of 15, i.e., 225. The standard deviations do not have to be the same for the two variables and they represent your best guesses about their SD values in your population. I discuss later ways you might "hedge" on these guesses.

I next need to specify the minimum important effect size that I want to be sensitive to in terms of the treatment effect on the mediator. This dictates both the value of the disturbance variance, d_1, and the value of the path coefficient for T→M, or p_1. Because T is a two valued dummy variable, p_1 is a mean difference that subtracts the mean M for the control group from the mean M for the intervention group. I specify the effect size using the 0 to 100 M metric and decide to set the minimal meaningful effect size of the program on study skills to be a mean difference of 7 points. This will be the value of p_1 in my simulation. I can then use this information to define the disturbance variance for M, var(d_1). Using the equation

$$var(M) = p_1^2\ var(T) + var(d_1)$$

I obtain

$$225 = (7^2)(0.25) + var(d_1)$$

and with algebraic manipulation I find that

$$var(d_1) = 225 - (7^2)(0.25) = 212.75$$

Note that using this value, I can calculate the proportion of unexplained variance in M; it is 212.75/225 = 0.946 and the proportion of explained variance in M by T is 1 – 0.946 = 0.054.

This exercise coupled with the previous steps yields the following population parameter values for the model thus far:

var(T)=0.25, var(M)=225, var(Y)=225, var(d_1)=212.75, and p_1=7.0

Next, I specify the minimal important effect size for the effect of study skills on exam performance, p_2. Given both M and Y are measured on a 0 to 100 metric, I might decide that a minimum path coefficient that I want to be sure to detect is 0.30 or greater for M→Y, i.e., for every one unit that I increase M, Y will increase, on average, 0.30 units. If the true path value is less than this, I judge the effect to be too small to be of interest and I do not care if I "miss it" via statistical significance testing in my study .

Once p_2 is specified, I derive the value of var(d_2). Using the equation

$$var(Y) = p_2^2\ var(M) + var\ (d_2)$$

I obtain

$$225 = (.3^2)(225) + var(d_2)$$

and with algebraic manipulation I obtain

$$var(d_2) = 225 - (.3^2)(225) = 204.75$$

The proportion of unexplained variance in Y when predicted from M is 204.75/225 = 0.910 and the proportion of explained variance (by M) is 1 – 0.910 = 0.090.

This gives me what I need for the simulation in terms of population parameters:

var(T)=0.25, var(M)=225, var(Y)=225, var(d_1)=212.75, var(d_2)=204.75, p_1=7.0, p_2=0.30

As you will see, I can determine the implications of altering these values by repeating the simulation on Mplus but using different parameter values but I show you first how to program Mplus using these particular parameter instantiations. I start by exploring the viability of using a total sample size of N = 150 (75 per group), which I might feel is pragmatically reasonable for the study I plan to conduct.

Stepping back for a big picture view, my goal in the simulation is to estimate statistical power under the analytic scenario of robust maximum likelihood as applied to the model in Figure 28.1. I seek to determine for N = 150 (a) the approximate statistical power for the link between T→M where the true population coefficient maps onto 5.4% explained variance, (b) the approximate statistical power for the link between M→Y where M accounts for 9.0% of the variance in Y in the population, and (c) as an ancillary analysis, the approximate statistical power for the omnibus mediation effect of (p_1)(p_2) = (7.0)(0.30) = 2.10, the value of which is dictated by my values of p_1 and p_2. As I will show you, there is much more information than this that I will be able to extract from the simulation.

Executing the Simulation

Table 28.1 presents the Mplus syntax to conduct the simulation.

Table 28.1: Local Simulation 1

```
1. TITLE: LOCAL SIMULATION 1;
2. MONTECARLO:
3. NAMES ARE t m y ;
4. CUTPOINTS = t(0);
5. NOBS = 150 ;          !sample size
6. NREPS = 20000 ;       !number of replicates
7. SEED = 2222 ;         !random seed
8. !SAVE = temp.dat;
9. ANALYSIS:
10. ESTIMATOR = MLR ;
11. MODEL POPULATION:    !specify population model
12. [t*0] ;              !set mean when generating original continuous t
13. t*1 ;                !set var when generating original continuous t
14. [y*0]; [m*0];        !set intercepts to 0
15. y ON m*.30 ;         !set effect of m on y
16. m ON t*7.0 ;         !set effect of t on M
17. y*204.75 ;           !disturbance variance for y
18. m*212.75 ;           !disturbance variance for m
19. MODEL:               !specify analysis model
20. y ON m*.30 ;         !outcome equation
21. m ON t*7.0 ;         !mediation equation
22. y*204.75 ;           !disturbance variance for y
23. m*212.75 ;           !disturbance variance for m
24. MODEL INDIRECT:
25. y IND t ;            !evaluate omnibus mediation effect
26. OUTPUT: TECH9 ;
```

Line 2 tells Mplus to conduct a simulation. Line 3 gives the names of the variables for which to generate data. Note there are no input variables from a data file. This is because Mplus generates the data internally using the computer. By default, Mplus creates scores for variables that are continuous and normally distributed. I need to dichotomize the treatment variable and I use Line 4 to do so. When Mplus encounters the `CUTPOINTS` subcommand it knows to break the continuous variables listed on the command, in this case `t`, into groups. The value in parentheses tells Mplus the cutpoint(s) to use to split the named variable into groups. If one value is listed there is one cutpoint. If two values are listed, there are two cutpoints. And so on. Line 4 indicates that the continuous variable `t` is to be converted to a binary variable because there is only one cutpoint specified. Any Mplus initially generated score for `t` that is less than 0 will be set to 0 and any initially generated

score greater than 0 will be set to 1. Given the initially generated t is, by default, normally distributed with a mean of 0, the command will result in t being transformed into two groups with an equal population probability of being in a given group. This results in two equal sized groups, scored 0 and 1.

Line 5 indicates the focal sample size I want to use, N = 150, or 75 per group. Line 6 is the number of replication studies you want Mplus to conduct. I indicate 20,000 of them. Line 7 provides a random number seed so that you can replicate the simulation results upon a second execution of the program. I discuss Line 8 later, which is commented out here. Line 9 tells Mplus that analysis commands and subcommands will follow. Line 10 tells Mplus to use a robust maximum likelihood estimator, and Line 11 states that you will next provide the population parameter values that I calculated earlier. Lines 12 to 18 provide traditional Mplus syntax to specify the parameter values by specifying the parameter name followed by a * and then the value of the parameter. Some people use a @ instead of a * when specifying population values in a Monte Carlo simulation. In Mplus, there is no difference between the two demarcations when you specify population parameters values in a simulation (but this is not so in normal Mplus programming or in the analytic phase of the simulation study).

Line 19 tells Mplus you will next specify the analytic model to apply in each sample. Lines 20 through 25 use Mplus syntax to specify the analytic model. These lines usually will be the same as those for the population model (lines 12 to 18). Note, however, that I provide starting values after the * sign that map onto the true population values specified in Lines 12 to 18. The one exception is that I omit reference to the exogenous variable t because Mplus by default treats exogenous variables as fixed predictors and I do not want t treated otherwise. This typically will be the case for all the exogenous variables in your simulation model; you will not reference their variances or covariances in the analysis portion of the simulation so they are treated as fixed predictors. The OUTPUT line invokes TECH9 which shows error messages that occur in the analysis of each simulation replicate.

Double Checking the Parameter Values

I make it a habit of double checking the population values I derived and specified in the syntax to ensure they reflect the dynamics I intended to produce. I do so using the syntax in Table 28.1 but I change Lines 5, 6 and 8 to read

```
5. NOBS = 2000000 ;
6. NREPS = 1 ;
8. SAVE = temp.dat;
```

and I run the revised syntax first. Mplus will then generate a single sample from the

population data consisting of an N = 2,000,000 cases and it stores the data in the file called temp.dat in the same folder that I ran the syntax in. Such a large N should produce results that are very close to the true population values. I can then analyze the data for this very large sample using a traditional Mplus program and double check my derivations. When I execute the above program first, I need to examine the Mplus output to see the order in which the variables were saved in the temp.dat file, like this:

```
SAVEDATA INFORMATION

  Order of variables

    M
    Y
    T
```

The order in which the three variables are stored in the temp.dat file is `M`, `Y`, and `T` in free format. I next run a standard Mplus model that maps onto the analysis model in Lines 19 to 25 of Table 28.1 as applied to this data set. Here is the syntax I use:

Table 28.2: Population Check

```
1. TITLE: POPULATION CHECK ;
2. DATA: FILE IS temp.dat ;
3. VARIABLE:
4. NAMES ARE m y t ;
5. ANALYSIS:
6. ESTIMATOR = MLR ;
7. MODEL:              !specify analysis model
8.    y ON m  ;        !outcome equation
9.    m ON t ;         !mediation equation
10.   y* ;             !disturbance variance for y
11.   m* ;             !disturbance variance for m
12. MODEL INDIRECT:
13.   y IND t ;
14. OUTPUT: SAMP STDYX TECH1 ;
```

All of the above syntax should be familiar to you.

Looking at the output for this program, I first double check the unstandardized population values to make sure they were what I specified. Here is the output for them (note: I ignore the standard errors and significance tests because they are meaningless byproducts of my true goals for the analysis, which is just to double check the population values):

```
                   Estimate       S.E.     Est./S.E. P-Value

 Y        ON
    M                 0.300      0.001    444.818      0.000

 M        ON
    T                 7.016      0.021    340.110      0.000

 Residual Variances
    M               212.795      0.213    999.949      0.000
    Y               205.124      0.205   1000.220      0.000
```

All of the values were what I programmed but with minor disparities that reflect sampling error, even with an N of 2,000,000. I calculated that p_1 should represent 5.4% explained variance in M as a function of T and that p_2 should represent 9.0% explained variance in Y as a function of M. I check the squared Rs on the output to see if this is, in fact, the case:

```
STDYX Standardization

R-SQUARE

    Observed                                        Two-Tailed
    Variable        Estimate       S.E.     Est./S.E.  P-Value

    M                  0.055      0.000    177.423      0.000
    Y                  0.090      0.000    233.227      0.000
```

All seems to be in order given I expect small disparities due to minor sampling error. Although the checking process is straightforward and may not seem necessary, when you start working with more complex models, this step can be helpful.

Output for Global Fit Indices

When we are certain that all is in order, we then execute the syntax in Table 28.1 to conduct the formal simulation. I now go over the output that results, beginning with the initial output focused on the global fit indices. Here is the output for the chi square statistic of global fit for the simulation:

```
Chi-Square Test of Model Fit

        Degrees of freedom                     1

        Mean                                1.029
        Std Dev                             1.455
        Number of successful computations   20000
```

Proportions		Percentiles	
Expected	Observed	Expected	Observed
0.990	0.990	0.000	0.000
0.980	0.980	0.001	0.001
0.950	0.951	0.004	0.004
0.900	0.900	0.016	0.016
0.800	0.800	0.064	0.064
0.700	0.704	0.148	0.153
0.500	0.505	0.455	0.466
0.300	0.309	1.074	1.114
0.200	0.209	1.642	1.703
0.100	0.106	2.706	2.774
0.050	0.052	3.841	3.902
0.020	0.022	5.412	5.567
0.010	0.010	6.635	6.726
0.020	0.021	5.412	5.564
0.010	0.011	6.635	6.712

The first part of the output reports the average chi square value across the 20,000 replications, which was 1.029. If the fit statistic truly has a sampling distribution that is chi square distributed, this value should roughly equal the model degrees of freedom, which is 1. In a set of scores that are chi square distributed, it turns out that the standard deviation of the chi square values will equal the square root of double the degrees of freedom. The square root of 2 is 1.414, which roughly equals the reported standard deviation value of 1.455 on the output. These results support the use of chi square p values to evaluate model fit for your study because the simulation suggests the chi square statistic has a sampling distribution that is chi square distributed, at least as reflected by the above two criteria.

The entry `Number of successful computations` indicates on how many of the simulation replications the solution converged. When sample sizes are small, non-convergence can be an issue and the result allows you to diagnose if this is a problem.

The column `Proportions Expected` is used in conjunction with the column `Percentiles Expected` to further evaluate if the sampling distribution of the fit statistic is chi square in nature. Each value in the `Proportions Expected` column provides the probability of observing a chi-square value greater than the corresponding value in the `Percentiles Expected` column for the operative degrees of freedom. For example, the proportion of 0.05 in the `Proportions Expected` column is the probability that the chi-square value exceeds the `Percentiles Expected` value (the critical value of the chi-square) of 3.841. The columns labeled `Observed` give the corresponding values calculated across the 20,000 replications. In this example, the observed probability 0.052 is close to the expected theoretical value of 0.050 and the observed percentile of 3.902 is close to the expected theoretical percentile of 3.841. These results again support that the chi square

statistic calculated under the study conditions are, in fact, chi square distributed.

Recall from earlier that the chi square statistic is based in asymptotic theory; only if the sample size is "large enough" will the sampling distribution of the statistic be chi square distributed and yield correct p values. The above simulation results affirm that for the model I am testing, my planned sample size of 150 is indeed "large enough" in terms of asymptotic theory. The simulation strategy represents an advantage over canned power analysis software for sample size selection because it allows us to evaluate sample size viability for asymptotic theory taking into account the number of variables, model complexity, and the patterning and magnitudes of the population covariances.

Here is the output for the CFI global fit index:

```
Mean                                  0.977
Std Dev                               0.057
Number of successful computations     20000

Cumulative Distribution Function
   Value    Function Value
   0.990       0.276
   0.980       0.238
   0.950       0.154
   0.900       0.078
   0.800       0.022
   0.700       0.008
   0.500       0.001
   0.300       0.000
   0.200       0.000
   0.100       0.000
   0.050       0.000
   0.020       0.000
   0.010       0.000
```

The "typical" or mean CFI across the 20,000 replicates was 0.977 with a standard deviation of 0.057. The output includes the cumulative distribution function (CDF) for the statistic. In this case, the CDF function value evaluated at the score X is the estimated probability that the CFI will take on a value less than or equal to X. For example, it is estimated that 15.4 percent of the CFIs in the sampling distribution of the CFI are less than or equal to 0.95 and that 100 – 15.4 = 84.6 percent of the CFIs are larger than 0.95. The vast majority of CFIs in this case were larger than 0.95. About 7.8% of the samples produced CFIs of 0.90 or less given the operative sampling error. Mplus reports this information for each of the global fit indices produced by Mplus. The often used rule of thumb of rejecting a model that has a CFI less than 0.95 seems, in this case, over-prone to rejecting a true model; 15.4% of the simulation trials have CFIs less than 0.95 even though the tested model is correct.

Output for Model Parameters

Here is the output for the core individual parameters in the model:

```
                                ESTIMATES                S. E.    M. S. E.    95%  % Sig
                   Population   Average   Std. Dev.     Average               Cover  Coeff

 Y        ON
  M                   0.300     0.3001     0.0789      0.0772     0.0062   0.939  0.966

 M        ON
  T                   7.000     7.0181     2.3648      2.3693     5.5924   0.948  0.842

 Residual Variances
  M                 212.750   209.9037    24.4369     23.8862   605.2332   0.924  1.000
  Y                 204.750   202.3009    23.7784     23.0077   571.3833   0.923  1.000
```

The column `Population` has the population value I specified for each parameter; the column `ESTIMATES Average` reports the average of the parameter estimates across the 20,000 replications. The values in these two columns should be close to one another. In this case, for the p_2 coefficient of M→Y, one value is 0.300 and the other is 0.3001. If they are close, this is evidence the parameter estimate is unbiased. Some researchers formalize the magnitude of bias by subtracting the population parameter value from the average parameter value, divide this number by the population parameter value, and then multiply the result by 100. For the effect T→M as captured by the regression of M on T, the bias is

[(7.0181 - 7.000) / 7.000] (100) = 0.259%

which is well below 1%. For the effect M→Y, the bias is

[(0.3001 - 0.300) / 3.000] (100) = 0.002%

which also is well below 1%. Standards vary for what is considered trivial bias; some methodologists suggest bias of 2.5% or less is acceptable (Bradley, 1978), but others set a standard of 5% or less, and still others suggest 10% or less is acceptable, although the criteria vary by context (Harwell, 2020).

The column `ESTIMATES Std. Dev.` is the standard deviation of the parameter estimate across the 20,000 replications. For the effect of T on M, it equaled 2.3648 and can be interpreted as the true standard error of population parameter in question. The column labeled `S.E. Average` is the average of the estimated standard errors calculated across the 20,000 simulation replications. The value of this average was 2.3693 for the effect of T on M. The discrepancy between it and the true standard error of the population parameter reflects the degree of bias in the estimated standard errors:

[(2.3693 - 2.3648) / 2.3648] (100) = 0.190%

which, again, is small and within ±1% of the true standard error.

If one doubles the true standard error, one obtains a sense of the margin of error that the sample size will yield for the parameter in question based on a 95% confidence interval. For the effect of T on M, the standard error on the output is 2.3648 so the estimate of the MOE that likely will result for N = 150 is about (2)(2.3648) = ±4.73. This estimate ignores assurance/tolerance probabilities. As such, across repeated samples about 50% of the MOEs will be lower than this value and 50% will be higher than this value. However, on average, the MOE for N = 150 will be about ±4.73. We thus garner some sense from the simulation about the MOE the sample size will produce.[2]

The column labeled `M.S.E.` is the mean square error for each parameter. It is the variance of the estimates across the 20,000 replications plus the square of the bias defined above. The column labeled `95% Cover` is the proportion of simulation replicates for which the 95% confidence interval contained the true population parameter value. This should be near 0.95. For the effect of T on M, it equaled 0.948. Note that if the value deviates too much in either direction from 0.95, this suggests potential problems. For example, if the 95% confidence interval coverage was close to 0.999, this suggests the estimated standard errors may be too large, making the intervals wider than they should be. Or, it might suggest the sampling distribution is ill-shaped. A common standard for 95% confidence interval coverage is that it should be within ±0.03 of 0.95 or between 0.92 and 0.98 to be satisfactory, although this standard can be adjusted upward or downward depending on context. Finally, the `% Sig Coeff` column is the proportion of times across the 20,000 replications that the null hypothesis was rejected. For a non-zero population value, it is the power of the statistical test. For a sample size of 150, the estimated power for the effect of T on M is 0.842 and for the effect of M on Y it is 0.966.

The power analysis simulation is based on a statistical test that uses robust maximum likelihood estimation grounded in asymptotic theory vis-à-vis SEM algorithms. In addition to statistical power, the simulation provides perspectives on bias of the parameter estimates, bias of the estimated standard errors of the parameters, the magnitude of margins of error, and confidence interval coverage, which is a strength of the approach.

The output also provides information for the omnibus indirect test in the model:

[2] The calculation also requires the estimator to be relatively unbiased and the sampling distribution to be normal in form. This may not be the case for variances, standardized coefficients, and indirect effects that are products of path coefficients. Also, some methodologists prefer to use the `S.E.Average` rather than the true standard error in the calculation. If the values are close, which is often the case, the choice is moot.

```
TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

                          ESTIMATES                 S. E.    M. S. E. 95%   % Sig
                Population   Average   Std. Dev.    Average           Cover   Coeff

Effects from T to Y

  Total              2.100     2.1071     0.9181     0.9102    0.8429  0.928  0.675
  Tot indirect       2.100     2.1071     0.9181     0.9102    0.8429  0.928  0.675

 Specific indirect 1
  Y
  M
  T                  2.100     2.1071     0.9181     0.9102    0.8429  0.928 0.675
```

The population value of the indirect effect T→M→Y is 2.10, which is the product of 7.00 times 0.30 or $(p_1)(p_2)$. The average of the parameter estimates (`ESTIMATES Average`) was 2.1071 suggesting the estimator for the indirect effect is unbiased in this case. The average estimated standard error for the indirect effect was 0.9102, which is close to the true population value of 0.9181, indicating it also is unbiased. The margin of error for the indirect effect will typically equal about (2)(0.91) = 1.82, but this should be considered a back-of-the-envelope descriptor. The confidence interval coverage is a bit low, but within the ±0.03 standard, namely it is 0.93. The statistical power for the omnibus indirect effect from T→M→Y was 0.675.

As discussed in Chapter 9, some methodologists argue that using bootstrapping for this test yields greater statistical power than the MLR method used in the simulation. I changed Line 10 to conduct bootstrapping in the simulation to the following:

```
ESTIMATOR = ML ; boot=2000  ;
```

and I added `CINTERVAL(BOOTSTRAP)` to the `OUTPUT` command. The power for the omnibus meditation test using bootstrapping changed upward to 0.782, with the values for power for the other parameter estimates staying about the same as they were in the MLR analysis. In this case, bootstrapping for the omnibus indirect effect was beneficial relative to increasing statistical power.[3]

I argued in previous chapters that the omnibus test of mediation often is of secondary interest in RET-based program evaluation. Given adequately powered tests of individual links in the mediational chain, there is useful information to be gained by analyzing the individual links even if the omnibus test is not adequately powered. To be sure, if I can pragmatically increase my sample size to obtain adequate power for the omnibus indirect test, then by all means, I do so. But I do not see power for the omnibus test as being essential

[3] Bootstrapping in simulations is computationally intense and can take considerable processing time.

to many of my program evaluations. Mind you, I would rather the test be sufficiently powered; but there are scenarios where practical constraints may not permit it. I re-ran the simulation with larger sample sizes using a trial-and-error process described below and found that a total sample size of approximately 180 (or 90 per group) is needed to bring the statistical power of the omnibus indirect effect test vis-à-vis the product of coefficients method to 0.80. In the document on my webpage called *Simulation Variants in Mplus*, I show you the Mplus code to calculate the statistical power of the joint significance test for the omnibus indirect effect. For N = 150, it equaled 0.814, which is somewhat higher than the product coefficient method, a result that is not atypical.

In sum, you can see that the local simulation strategy yields a wealth of information relevant to sample size selection compared to commercially available canned software for power analysis. I provides information about the viability of asymptotic theory, the accuracy of global fit indices, the use of a robust estimator, parameter bias, standard error bias, margins of error, confidence interval coverage and statistical power for every parameter in the model including direct and indirect effects. It should be your method of choice for sample size evaluation.

Statistical Power for the Global Chi Square Test

In the current simulation, the model I tested is correctly specified relative to the population model. This means that statistical power for the global chi square test of fit cannot be evaluated for it. It is possible to conduct power analysis for the global chi square test using a simulation strategy in which you *a priori* posit specification error for a misspecified model and determine how sensitive the chi square test is to detecting the misspecification. I describe a strategy for doing so in the document *Simulation Variants in Mplus* on my website. With only one degree of freedom for the current model, the most theoretically coherent source of specification error is one that omits the dashed path *c* in Figure 28.1 when, in the population, that path contributes to the outcome. In the current simulation this path was assumed to be zero and omitted from the model I tested. If I set the true strength of path *c* in the population to be 7.00 (the same as T→M) and I fit a misspecified model that ignores this path, what is the power I will have for an N of 150 to reject the misspecified model vis-à-vis the chi square test of fit? Using the methods discussed in *Simulation Variants in Mplus*, I found that it is approximately 0.86.

Exploring Sample Sizes, Effect Sizes, and Model Parameter Values

I usually conduct my initial simulation using as a reference the largest sample size I think I can pragmatically muster in my main study. I then explore smaller sample sizes in the simulation to see the implications of using smaller sample sizes. Also, as noted, I like to

think about power analyses using the concept of effect size sensitivity. For example, for the intervention to increase study skills using a sample size of 150, what is the effect size sensitivity for the effect of the intervention on study skills assuming power of 0.80 for an alpha level of 0.05 with a two tailed test. In the simulation reported above, the power for T→M when p_1 equaled 7.0 was 0.842. To document effect size sensitivity, I need to re-run the simulation but lower the value of p_1 in successive runs until the statistical power associated with the lowered p_1 value is 0.80. Here are the original population parameters I used in the simulation:

var(T)=0.25, var(M)=225, var(Y)=225, var(d_1)=212.75, var(d_2) = 204.75, p_1=7.0, p_2 = 0.30

I know I am close to having statistical power of 0.80 when p_1 = 7.0, so I decide to rerun the simulation anew by setting p_1 = 6.5. This, in turn, will affect var(d_1). Recall that I calculated var(d_1) as follows:

$$var(M) = p_1^2\, var(T) + var(d_1)$$

$$225 = (7^2)(0.25) + var(d_1)$$

$$var(d_1) = 225 - (7^2)(0.25) = 212.75$$

With p_1 = 6.5, this becomes

$$var(d_1) = 225 - (6.5^2)(0.25) = 214.438.$$

Here is the original syntax but with these new values:

Table 28.3: Revised Syntax 1

```
1. TITLE: LOCAL SIMULATION 1;
2. MONTECARLO:
3. NAMES ARE t m y ;
4. CUTPOINTS = t(0);
5. NOBS = 150 ;          !sample size
6. NREPS = 20000 ;       !number of replicates
7. SEED = 2222 ;         !random seed
8. !SAVE = temp.dat;
9. ANALYSIS:
10. ESTIMATOR = MLR ;
11. MODEL POPULATION:    !specify population model
12. [t*0] ;              !set mean when generating original continuous t
13. t*1 ;                !set var when generating original continuous t
14. [y*0]; [m*0];        !set intercepts to 0
15. y ON m*.30 ;         !set effect of m on y
```

```
16. m ON t*6.5 ;         !set effect of t on M
17. y*204.75 ;           !disturbance variance for y
18. m*214.438 ;          !disturbance variance for m
19. MODEL:               !specify analysis model
20. y ON m*.30 ;         !outcome equation
21. m ON t*6.5 ;         !mediation equation
22. y*204.75 ;           !disturbance variance for y
23. m*214.438 ;          !disturbance variance for m
24. MODEL INDIRECT:
25. y IND t ;            !evaluate omnibus mediation effect
26. OUTPUT: TECH9 ;
```

Here is the output for just the effect of T on M:

```
                                ESTIMATES             S. E.  M. S. E.   95%  % Sig
                  Population     Average   Std. Dev.  Average           Cover Coeff

 M        ON
  T                   6.500     6.4862     2.4223     2.3722   5.8649  0.942 0.777
```

The power is 0.777 when $p_1 = 6.5$. Given statistical power was 0.842 when $p_1 = 7.0$, my next educated guess is that I will probably obtain power of 0.800 (give or take) for the effect of T→M when $p_1 = 6.7$. When I re-ran the simulation making the requisite changes, the estimated power was 0.805, affirming that the effect size sensitivity for an N = 150 for T→M is approximately 6.70 or greater.

My general point is that you can address most any question related to fluctuations in sample size, power, bias, asymptotic theory, margins of error, and confidence intervals by varying different parameters and/or sample sizes within the localized simulation. Experimenting with different sample sizes is particularly easy because it does not require changing any other parameters in the simulation.

Checking Type I Error Rates

My analyses to this point have focused on issues surrounding statistical power, but it also can be helpful to determine how your planned study conditions affect Type I error rates. As an example, suppose the true effect of the study skills on exam performance is zero, i.e., p_2 is zero. Using an alpha level of 0.05, the proportion of times I should reject the null hypothesis when estimating the effect M→Y in my model should be close to 0.05. I find it helpful to verify that my planned sample size also will perform satisfactorily for Type I error rates, especially when I am using a statistical method that relies on asymptotic theory. Here are the population parameter values I would use in this case based on my original simulation but now where $p_1 = 7.0$ and $p_2 = 0$:

var(T)=0.25, var(M)=225, var(Y)=225, var(d_1)=212.75, var(d_2)=225, p_1=7.0, p_2=0

Note that var(d_2) has changed to equal var(Y) because M explains no variation in Y. I made the corresponding changes in the syntax and here are the results for M→Y:

		ESTIMATES			S. E.	M. S. E.	95%	% Sig
		Population	Average	Std. Dev.	Average		Cover	Coeff
Y	ON							
M		0.000	-0.0010	0.0825	0.0811	0.0068	0.942	0.058

The proportion of incorrect null hypothesis rejection was 0.058, which is reasonably close to 0.05. This also was true for p_1 when I set p_1 to 0 and p_2 to 0.30. The Type I error rates seem to be in order for the simulation conditions I evaluated.

Programming More Complex Models

The model I used to develop the logic of simulations was unrealistic in that it consists of a single mediator and a single outcome with no covariates. The RET models you work with likely will be more complex with multiple mediators and multiple covariates. Setting up such simulations can be challenging. Here, I give you an appreciation for relevant issues by augmenting the model in Figure 28.1 to include covariates and a direct effect from the treatment to the outcome, per Figure 28.2. The covariates are each measured on a 0 to 100 metric and take the role of baseline mediator/outcome measures. The baseline exam performance reflects how the student performed on the prior math exam s/he took. The figure omits correlations between exogenous variables to reduce clutter, but they are taken into account.

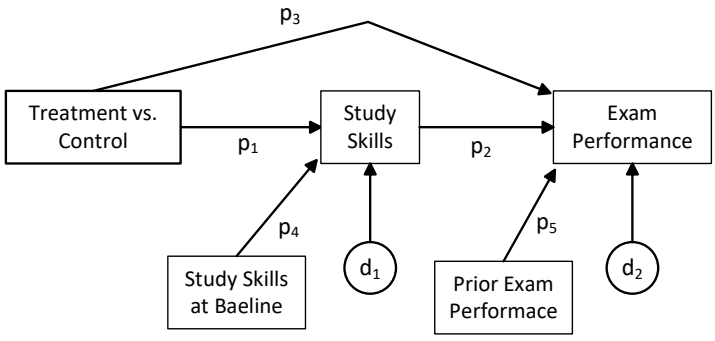


FIGURE 28.2. Simulation Example with Single Mediator and Covariates

Here are the two equations that capture this model, per the discussion of model equations in Chapter 7:

$$SS = a_1 + p_1\ T + p_4\ SSB + d_1 \quad [28.9]$$

$$EP = a_2 + p_3\ T + p_2\ SS + p_5\ EPB + d_2 \quad [28.10]$$

where SS = study skills at the posttest, SSB = study skills at baseline, EP = exam performance at the posttest, EPB = exam performance at baseline, and T is the treatment dummy variable.

For Equation 28.9, I need to specify for the simulation the variance of the criterion (SS), the variance of each of the predictors, the path coefficients p_1 and p_4, and the variance of the disturbance term, d_1. For Equation 28.10, I need to specify the variance of the outcome (EP), the variance for each of the predictors, the path coefficients for p_2, p_3, and p_5, and the variance of the disturbance term. As before the intercepts can be fixed to zero (i.e., ignored) as they do not enter into model analyses. In the original model in Figure 28.1, an equation I used to help me assign values was Equation 28.6 that described the relationships between variances in a linear equation $Y = \alpha + \beta\ X + \varepsilon$ as follows:

$$var(Y) = \beta^2\ var(X) + var\ (\varepsilon)$$

It turns out that when there is more than one predictor in the equation, the function for the variances is more complex. For the case of two predictors, X1 and X2, of an outcome Y, the linear equation is $Y = \alpha + \beta_1\ X_1 + \beta_2\ X_2 + \varepsilon$ and the variance equation becomes:

$$var(Y) = [\beta_1^2\ var(X_1) + \beta_2^2\ var(X_2)] + [(2)(\beta_1)(\beta_2)cov(X_1,X_2)] + var(\varepsilon) \quad [28.11]$$

The first expression in brackets on the right side of the equation sums the square of the path/regression coefficients for each predictor times the variance of the predictor. The last term on the right hand side of the equation is the disturbance variance, var(ε). The middle bracketed term takes into account the covariances between the predictors, also known as collinearity. In a single predictor model, the middle term drops out because there is no collinearity. When working with a linear equation with more than one predictor, we must take it into account.

I can write the above variance function using more formal statistical notation for the general case of *k* predictors and using population notation as follows:

$$\sigma_Y^2 = [\sum_{j=1}^{k} \beta_j^2 \sigma_j^2] + [2\sum_{m=1}^{k}\sum_{j>1}^{k} \beta_m \beta_j\, cov(X_m, X_j)] + \sigma_\varepsilon^2 \quad [28.12]$$

Equation 28.12 may appear intimidating but there are simple workarounds for using it to formulate your simulation. Because the details are somewhat cumbersome, I do not develop them here. They are described in the document *Simulation Variants in Mplus* on my webpage.

At this point, I have described the fundamental concepts for conducting a localized simulation for power analysis using Mplus. I apply these concepts to many different scenarios in the document *Simulation Variants in Mplus* on my webpage. I urge you to look at that document to help you master how to conduct localized simulations. In the next sections, I show you how to conduct a localized simulation using an existing data set, which I often find to be helpful. I then describe how to discuss sample size simulations with clients and incorporate them into publications and grant proposals. The final section of this chapter is titled *Small Sample Statistical Tests*. It provides guidance on how to approach analytics when you are faced with small sample sizes and must use specialized small sample tests.

Post Hoc Localized Simulations

An exercise I sometimes find helpful is to perform a localized post hoc simulation using the data I have collected in my main study but only after I feel I have isolated what I believe is a satisfactory explanatory model for the data, i.e., I have isolated my "final" SEM model. As noted earlier, post hoc power analysis is generally frowned upon, but my goal in conducting this simulation is not to evaluate power but instead to increase the confidence I have in the analytics I have used (e.g., the applicability of asymptotic theory, adequate confidence interval coverage). The analysis is simple to implement and often better maps onto the realities of the data than a traditional simulation.

Suppose I collect data for an N of 130 individuals for an RET study that randomly assigns individuals to an intervention versus control condition (about 65 per group). Suppose I have three mediators and a single outcome. All the variables except the treatment condition are measured on muti-item scales with total scores ranging from 0 to 15 in the form of discrete integers. The treatment condition is dummy coded, 0 = control group, 1 = intervention group. Figure 28.3 presents the model I used for my main analysis. The model does not include covariates to make my exposition simpler but the presence of covariates will be typical. Suppose I posited the model a priori and it ultimately provided good fit to the data. Based on the data and prior theory, I believe the model represents a reasonable accounting of the causal dynamics at play. The figure also presents the unstandardized coefficients for key parameters in the fitted model to the data. The intercepts for each endogenous variable are shown in brackets beneath the respective endogenous variable. The disturbance terms for the three mediators are correlated with one another to account for unmeasured common causes of them.

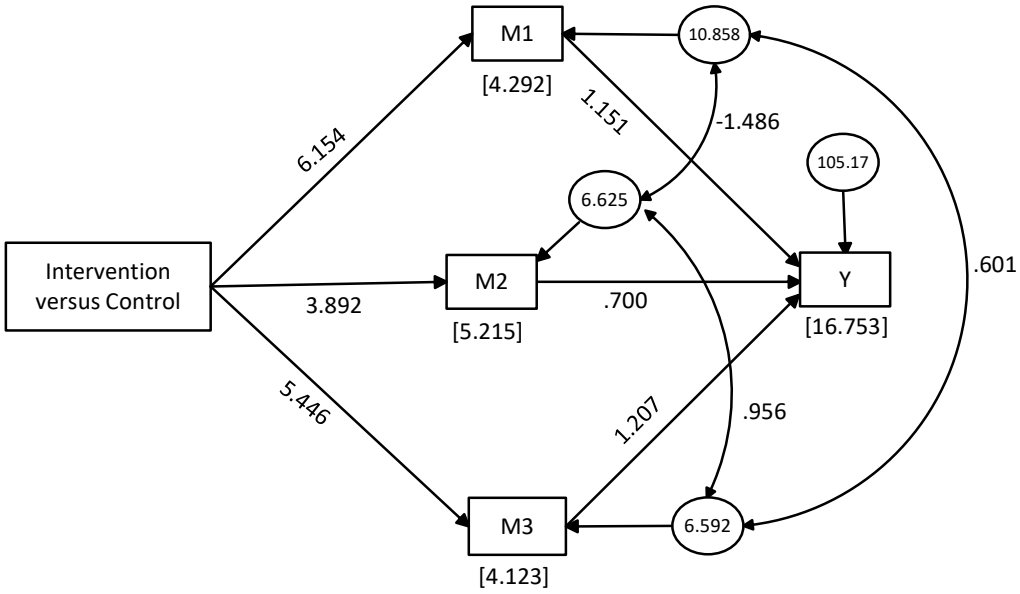


FIGURE 28.3. Post hoc simulation example

For my SEM data analysis, I might be concerned that my sample size (N = 130) is too small to satisfy the requirements of asymptotic theory. When I conducted analyses on the data, I also noted that M1 had a distinctly non-normal distribution that was close to being uniform in nature. Although I used a robust estimator (MLR) in my analysis, I am unsure if the robust estimator can overcome this type and amount of non-normality given the somewhat small sample size. Finally, each of the endogenous measures is scored from 0 to 15 using discrete integers. Technically, these are not continuous measures. A continuous measure has an infinite (or large) number of possible points between any two locations on the scale (e.g., a measure of time, in theory, has an infinite number of values between 1 and 2 seconds). Perhaps my measures are too crude for SEM to function reasonably (although prior research reviewed in Chapter 3 suggests that using measures with 7 or more discriminations should be adequate). I seek reassurance that my model and analysis is reasonable under the simultaneous operation of all of these conditions and I do so using the specialized post hoc simulation method. The method assumes the target model, in this case the one in Figure 28.3, is correctly specified and reasonably captures the underlying causal dynamics, which I believe is the case here. The population model is not assumed to fit the data perfectly, i.e., it does not necessarily produce a zero residual matrix of all zeros. But the model is deemed to be a "close enough" fit in the population to be descriptive of the core causal dynamics.

To execute the simulation, I first treat the data I have collected as population data and I generate 1,000 random samples of size N = 130 from it by using *sampling with replacement* from my sample data, much like we do in bootstrapping. By using sampling with replacement, I essentially treat my collected data as population data coming from a population that is infinite in size but that results from a causal model having parameter values equal to those observed in Figure 28.3. I generate the 1,000 samples using the bootstrapping option in Mplus but as a data generation tool rather than as a means of using bootstrapping to estimate standard errors, p values, and confidence intervals. In other words, I ignore the results of the bootstrap per se but instead use it to generate data samples for eventual use in Mplus' external simulation option. Table 28.4 presents the Mplus code I used to generate the samples.

Table 28.4 Syntax for Post Hoc Simulation

```
1.  TITLE: generate bootstrap samples ;
2.  DATA: FILE IS c:\mplus\posthoc.txt ;
3.  VARIABLE:
4.    NAMES ARE treat m1 m2 m3 y ;
5.    MISSING ARE ALL (-9999) ;
6.  ANALYSIS:
7.    ESTIMATOR = ML ; BOOTSTRAP = 1000 ;
8.  MODEL:
9.    y ON m1 m2 m3 ;
10.   m1 ON treat  ;
11.   m2 ON treat  ;
12.   m3 ON treat  ;
13.   m1 m2 m3 WITH  m1 m2 m3   ;
14. SAVEDATA: SAVE=BOOTSTRAP; FILE IS bootreps*.dat ;
15. OUTPUT: SAMP RESIDUAL STAND(STDY) CINTERVAL TECH4  ;
```

All of the syntax should be familiar except Lines 13 and 14. Line 13 is a shorthand way of telling Mplus to correlate all the variables on the left side of the `WITH` term with all the variables on the right side of it. In this case, because the mediators are endogenous, the syntax essentially tells Mplus to correlate all the disturbance terms of the mediators. Line 14 tells Mplus to save each bootstrap sample in a separate file and the names to give to those files. The naming convention is the same as that for generating imputed data sets that I described in Chapter 26. Each data set will be named "bootreps" from the `FILE IS` subcommand (you can use any name you want) followed by a number from 1 to 1,000 (the number of requested data sets) with the tag "dat" attached to each data set (you can use any tag designation you want). Mplus also will generate a file called "bootreps", which is the name you assigned to each data file but now the name will be followed by the word

"list.dat" instead of a number with the "dat" tag. This file contains the list of names of all the generated data sets in a single column for input into the program that ultimately conducts the simulation and which I show you next. All of the saved files will be in the same folder that the input syntax is stored in because I do not specify a folder path for them.

The data are written to the sample data files in an order that Mplus tells you at the end of the output from the run in Table 28.4:

```
SAVEDATA INFORMATION

Order of variables

    M1
    M2
    M3
    Y
    TREAT
```

Finally, Table 28.5 presents the syntax for the subsequent run that tells Mplus how to structure the simulation analysis and the population values to use, all taken from my final SEM analysis of the main data.

Table 28.5 Syntax for Final Step of Post Hoc Simulation

```
1.  TITLE: LOCAL SIMULATION ANALYSIS ;
2.  DATA: FILE = bootrepslist.dat ;
3.  TYPE = MONTECARLO ;
4.  LISTWISE = ON ;
5.  VARIABLE:
6.    NAMES ARE m1 m2 m3 y treat  ;
7.    MISSING = ALL(-9999);
8.  ANALYSIS:
9.  ESTIMATOR = MLR;
10. MODEL:
11.   y ON m1*1.151 m2*.700 m3*1.207 ;
12.   m1 ON treat*6.154 ;
13.   m2 ON treat*3.892 ;
14.   m3 ON treat*5.446 ;
15.   m1 WITH m2*-1.486 ;
16.   m1 WITH m3*.601 ;
17.   m2 WITH m3*.956 ;
18.   [m1*4.292] ;
19.   [m2*5.215] ;
20.   [m3*4.123] ;
21.   [y*16.753]  ;
22.   m1*10.858  ;
```

```
23.    m2*6.625 ;
24.    m3*6.592 ;
25.    y*105.170 ;
26. MODEL INDIRECT:
27.    y IND treat ;
28. OUTPUT:
```

Most of the syntax should again be familiar. Lines 11 to 25 specify the model and the "population" values for each parameter. Again, these values were taken from the final model I had settled upon when applying the model to the sample data (per the values in Figure 28.3).

The global chi square test is a test of the null hypothesis that the population residual matrix is all zeros. This is not the state of affairs in the population because the population fitted model was deemed "close enough" but not "perfect" and produced a population residual matrix that was near zero but not exactly zero. I therefore ignore the simulation results for the global chi square test per se.

Here are the means and standard deviations across the 1,000 simulations for the other global fit indices I tend to rely on:

```
RMSEA (Root Mean Square Error Of Approximation)

        Mean                                0.045
        Std Dev                             0.069

CFI

        Mean                                0.997
        Std Dev                             0.008

SRMR (Standardized Root Mean Square Residual)

        Mean                                0.008
        Std Dev                             0.006
```

The results for both the CFI and SRMR suggest these statistics are well behaved. The RMSEA mean is reasonable but the RMSEA standard deviation suggests more variability than ideal for this statistic.

Here are the simulation results for the model coefficients:

		ESTIMATES		S. E.	95%	% Sig
	Population	Average	Std. Dev.	Average	Cover	Coeff
Y ON						
M1	1.151	1.1525	0.2483	0.2331	0.929	0.993
M2	0.700	0.6972	0.3040	0.3020	0.954	0.633
M3	1.207	1.2023	0.3198	0.3098	0.936	0.964
M1 ON						
TREAT	6.154	6.1538	0.5885	0.5713	0.939	1.000
M2 ON						
TREAT	3.892	3.9056	0.4474	0.4493	0.957	1.000
M3 ON						
TREAT	5.446	5.4546	0.4448	0.4473	0.952	1.000
M1 WITH						
M2	-1.486	-1.4458	0.8068	0.8179	0.941	0.431
M3	0.601	0.5896	0.7032	0.6958	0.954	0.135
M2 WITH						
M3	0.956	0.9573	0.5414	0.5362	0.947	0.425
Intercepts						
M1	4.292	4.2866	0.3853	0.3599	0.923	1.000
M2	5.215	5.2008	0.2973	0.2921	0.946	1.000
M3	4.123	4.1153	0.3046	0.2932	0.932	1.000
Y	16.753	16.8547	2.4335	2.3998	0.944	1.000
Residual Variances						
M1	10.858	10.6234	2.2140	2.1885	0.914	1.000
M2	6.625	6.5292	0.7136	0.6912	0.925	1.000
M3	6.592	6.4641	0.7482	0.7290	0.928	1.000
Y	105.170	102.4545	13.4782	12.7649	0.914	1.000

I focus here on the results for the regression of `M3 ON TREAT` to highlight what I look for throughout the output for the other core model parameters. First, the population mean difference on `M3` as a function of `TREAT` was 5.446. The mean of the difference across the 1,000 samples was 5.4546, which is quite close to the "population" value. This suggests the sample mean difference estimates across the samples are functionally unbiased. The standard deviation of the regression coefficients when `M3` was regressed onto `TREAT` was 0.4448 (the standard error of the coefficient). The typical standard error in the 1,000 samples was 0.4473, which is quite close to the calculated standard error. This also is a desirable property for estimators. The coverage rate of the 95% confidence interval was 0.952, which is quite close to the theoretical expectation of 0.950. Statistical power is reported in the last column but this is not of interest because of the arbitrary nature of the operative meaningfully important effect sizes. Rather, I am more interested in performance

relative to parameter bias, bias in the parameter standard error, and the confidence interval coverage. These properties seem to be in order and this is true of all the other parameters in the model, with the possible exception of confidence interval coverage rates for some residual variances.

Here are the results for the analysis of the total and indirect effects using the same output format:

```
TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS


                          ESTIMATES              S. E.        95%      % Sig
                Population   Average   Std. Dev.   Average      Cover     Coeff

Effects from TREAT to Y

  Total             16.381    16.4106     1.6446     1.6268     0.955     1.000
  Tot indirect      16.381    16.4106     1.6446     1.6268     0.955     1.000

 Specific indirect 1
  Y
  M1
  TREAT              7.083     7.0889     1.6777     1.5652     0.921     0.993

 Specific indirect 2
  Y
  M2
  TREAT              2.724     2.7446     1.2647     1.2602     0.952     0.588

 Specific indirect 3
  Y
  M3
  TREAT              6.573     6.5770     1.9011     1.8411     0.942     0.959
```

The simulation results for this facet of the analysis also seem reasonable. Based on the above I feel better about the properties of my main analysis in terms of asymptotic theory, confidence interval coverage, bias, and margins of error as well as the global fit indices (excluding the chi square statistic) with the possible exception of the RMSEA.

In sum, I sometimes use the above post hoc simulation strategy to gain perspectives on the confidence I can have on facets of my modeling efforts. The approach is not without its limitations but I find it useful to alert me to potentially problematic issues when analyzing my data. The approach assumes the tested model lacks specification error or if specification error is present, it is inconsequential (see also an alternative approach described in West et al., 2023 and Wolf & McNeish, 2021, 2023).

Writing Up a Sample Size Decision Making Simulation

Many clients for whom I evaluate programs have little sense of the concepts of statistical power, margins of error, effect size sensitivity, confidence interval coverage, parameter estimation bias, and robustness. However, they are keenly interested in sample size choices because of the cost implications. I find it easiest to explain intuitively the concepts of margins of error and effect size sensitivity to clients. Indeed, talking through these concepts with them is important so I can determine their tolerances for margin of error magnitudes and effect size sensitivity. In my final reports for clients, I invariably include a technical appendix that describes the simulations I conducted and their results to justify sample size selection. However, this is more for the benefit of readers who are technically oriented.

For journal publications, most journals have page or word limits that do not allow me to describe my simulations in detail. However, most reputable journals maintain on-line supplements and I can provide the needed details there. My write ups typically include the parameter values that guided the simulation, justification for the selection of parameter values, and the simulation results. For the results, I emphasize what happened in terms of asymptotic theory, statistical power, margins of error, confidence interval coverage, Type I error rates, missing data, non-normality, and bootstrap viability, as appropriate.

Most grant proposals address statistical power but the common practice is to rely on canned power analysis software rather than simulations. Space constraints often restrict the presentation of power analysis details and other factors I might consider that drive my choice of a sample size. As analytics have become increasingly complex and as NIH now often adds a "quantitative expert" to review proposals who may lack substantive expertise in one's content area, it can be challenging to address sample size selection in the detail it deserves given space constraints. I find it helpful to frame sample size decisions in grant proposals using the effect size sensitivity framework in conjunction with margin of error analysis.

Concluding Comments on Localized Simulations

Localized simulations have many advantages. They force researchers to think through all of the assumptions they must make when evaluating statistical power and they address a much broader scope of sample size related issues than canned power analysis software. This includes statistical power, the applicability of asymptotic theory, margin of error analysis, parameter bias, confidence interval coverage, Type I error rates, robustness, missing data, the viability of bootstrapping, convergence issues, and covariance instability. Disadvantages are that they require acquiring simulation skills and they can be time consuming. I think having this tool in your analytic and design toolbox is a desirable goal.

SMALL SAMPLE STATISTICAL TESTS

At times, we must conduct RETs with small sample sizes because of financial, practical and/or logistical constraints. We still want to use quantitative tools from our toolbox but full information SEM (FISEM) may not be practical. One strategy in such scenarios is to use FISEM but with adaptations that can accommodate the smaller sample size. Another strategy is to move to limited information SEM (LISEM) that focuses on the separate equations of an SEM model but then apply well developed small sample methods (such as ordinary least squares or a robust regression counterpart of it) to each equation and then piece together the results of the individual equation analyses to form the larger model results. I discussed the strengths and weaknesses of FISEM versus LISEM in Chapter 8 and have provided multiple illustrations of the viability of LISEM throughout prior chapters, even when latent variables are involved. In this section, I first review small sample FISEM approaches you might consider and then I describe small sample statistical methods that can be used in LISEM. None of these approaches solve the problem of low power and poor effect size sensitivity that so often accompanies small sample sizes. However, as discussed earlier in this chapter, low power and poor effect size sensitivity often can be countered by the strategic use of covariates and design considerations, albeit with some constraints/limitations.

How do you know if you have a "small sample?" It turns out this is not a simple question to answer. I agree with McNeish (2017) that such characterizations of sample size are relative; it depends on the absolute sample size, model complexity, model type, and the metric properties of the variables in the model, among other things. A sample can be considered small if it does not contain sufficient information to satisfactorily evaluate one's model by virtue of the number of individuals in the sample. Judgments of "smallness" thus reflect the consideration of a range of factors, not just sample size per se.

Small sample size issues are related to what is known as the **curse of dimensionality** that I discussed in Chapter 17. The curse of dimensionality in the machine learning literature is when your data have too many features and you need to somehow reduce the number of features to make the situation workable and amenable to analysis. The current section does not focus on data reduction methods per se but on statistical methods that can be brought to bear to RET analyses that embrace the full RET design or to compartmentalizations of the RET design into sections each of which can be analyzed using small sample methods.

Small Sample Full Information SEM

Many simulations have shown that maximum likelihood estimation with small samples can result in convergence problems, inadmissible parameter solutions, and biased estimates (e.g., Boomsma, 1985; Nevitt & Hancock, 2004). In FISEM, a number of corrections to the traditional chi square test of model fit have been suggested to compensate for the fact that the statistic often is not chi square distributed when the sample size is small relative to model complexity. Empirical evaluations of these corrections have operationalized model complexity differently, such as by the number of observed variables in the model, the ratio of the sample size to the number of observed variables in the model, the number of estimated parameters in the model, and the model degrees of freedom. Herzog, Boomsma, and Reinecke (2007) and Herzog and Boomsma (2009) concluded that for complete data that is multivariately normally distributed, a chi square correction proposed by Swain works well as long as the sample size is at least 50 and the ratio of the sample size N to the number of estimated parameters is at least 2:1. If data are non-normal, Boomsma and Herzog (2013) argue the correction can be applied to the robust maximum likelihood chi squares of MLR and MLM in Mplus to help mitigate the adverse effects of nonnormality. Shi, Lee, and Terry (2018) recommend the use of a correction proposed by Yuan et al. (2015) when the number of observed variables is large (e.g., greater than 90) but they also affirm the satisfactory performance of the Swain index as long as the number of variables is less than 90 and the sample size is at least 50 with a 2:1 N to number of estimated parameters ratio.

McNeish (2020) recommends the use of an F statistic computed as the model chi square statistic divided by its degrees of freedom and found that it worked reasonably well with respect to the control of Type I error rates for sample sizes as small as 50 for a three factor, 15 variable CFA model. McNeish (2020) also suggests that one also can apply the F test to robust chi square statistics (e.g., MLM, MLR) rather than the just traditional chi square statistic.

The program *Small N SEM Corrections* on my website allows you to apply all the aforementioned corrections. The F test (as well as the other tests) assume no missing data in the sense they require the sample size as input into the formulae they apply. With missing data, it is unclear what the sample size should be defined as being for the formulae. McNeish and Harring (2017) suggest an ad hoc approach for defining the N in such scenarios, but not enough is known about its utility. Another option is to define the operative N multiple times across a range of N reflect by the amount of missing data and then evaluate the robustness of conclusions across the different sample size scenarios in the spirit of a sensitivity framework.

The above corrections apply to the global chi square test of fit but they do not generalize to the standard errors for significance tests of individual parameter estimates nor to modification indices. McNeish (2020) suggests using a *t* distribution to define critical values for these single degree of freedom null hypothesis tests, but it is unclear if this strategy is viable. The approach clearly is more conservative than relying on the traditional critical ratios in SEM output, but its potential use and limitations needs further exploration.

Another full information estimation approach for dealing with small sample sizes is to use Bayesian SEM (see Chapter 8). McNeish (2016b) as well as others (Smid, McNeish, Miočević & van de Schoot, 2020) argue that this strategy tends to work reasonably well with informative priors but if you use diffuse priors (which is common practice) with small sample sizes, the result can be non-trivial parameter bias, often as much or even more so than traditional SEM methods. Another problem is that with small sample sizes, the influence of the prior distribution on the posterior distribution can be substantial thereby overpowering the impact of the data per se on the conclusions one makes. Bayesian modeling is not a panacea for small sample size research despite claims you may sometimes encounter to the contrary. However, when coupled with well-conceived prior distributions, it can be useful.

One strategy for specifying an informative prior is to reduce the parameter space in question so that it does not consider ***impossible values***, i.e., values that are impossible do not receive any density mass in the prior distribution. Another approach is to restrict the parameter space from taking ***implausible values***, i.e., values that receive very small density mass in the prior distribution but that could, in theory, be obtained after the prior has been updated with the data. As examples, variance parameters cannot take on values less than zero, so one might include this restriction into the prior distribution of a variance. If a continuous mediator, M, has a metric that ranges from 1 to 7 and it is regressed onto a 0-1 dummy variable for the treatment condition, T, the intercept represents the mean on M for the control group, i.e., the mean on M when T = 0. This intercept value cannot be outside the range of 1 to 7, so the prior distribution for it can be specified to have these bounds.

In Mplus, prior distributions are specified using the `MODEL PRIORS` command, which is placed just after the `MODEL` command and before the `OUTPUT` command. I illustrate relevant syntax using an example where I regress the above mediator, M, onto the binary treatment condition, T:

```
1. MODEL:
2. M ON T ;
3. [M] (a) ;          ! specify intercept and give label
4. M (varM);          ! specify disturbance variance of M and give label
5. M ON T (p1) ;      ! specify path coefficient of M on T and give label
6. MODEL PRIORS :     ! assign informative priors you want
```

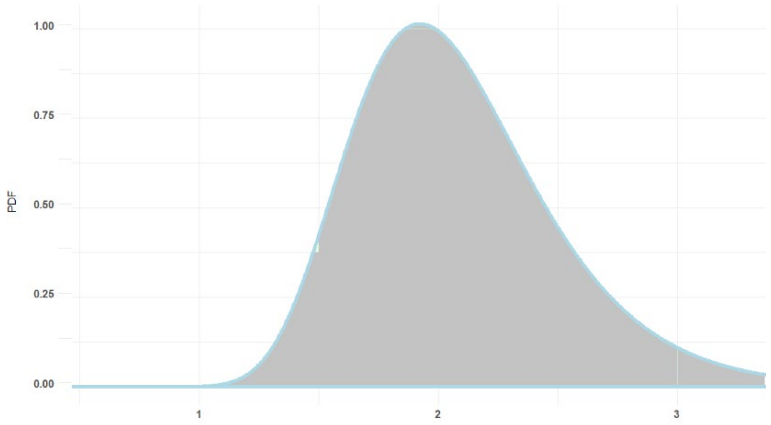
```
7.   a~U(1,7) ;
8.   p1~N(3, 1) ;
9.   varM~IG(25, 50);
10. OUTPUT: CINTERVAL(HPD) TECH8;
```

Line 6 tells Mplus you will specify one or more prior distributions. Line 7 sets the prior distribution for the intercept to a uniform distribution using the letter `U` preceded by a ~ symbol. The number in parentheses, 1 to 7, are the distribution boundaries; it excludes values outside the 1 to 7 boundaries. The narrower the range, the more informative the prior is. The path coefficient, p1, translates in this example to a mean difference between the treatment and control group. It is not uncommon to set an informative prior distribution for a regression coefficient to a normal distribution with a specified mean and variance. On Line 8, I indicate that the prior distribution of p1 should be a normal distribution (indicated by the letter N) with a mean of 3.0 and a variance of 1.0. In practice, I would specify the value of the prior distribution mean based on past research, meta-analyses, expert opinions and/or common sense and the value of the variance to reflect my uncertainty about the true mean value; the larger the variance, the more uncertain I am (see Chapter 8). If my model included a covariate, Z, in the equation, then the values from past research and meta-analyses from which I took the mean value for p1 must also take into account Z.

The prior distribution for a disturbance variance often is specified as an inverse gamma distribution, denoted in Mplus by `IG`. It has two hyperparameters, α and β, which are the shape and scale parameters of the distribution, respectively. Suppose you use a prior study or a pilot study to determine the values of these hyperparameters. Gelman et al. (2013; see also the discussion at https://stats.stackexchange.com/questions/350924/why-do-we-use-inverse-gamma-as-prior-on-variance-when-empirical-variance-is-gam) recommend setting α to half of the sample size of the previous study, and β to half the sample size of the previous study times the variance estimate from that study. Suppose based on a pilot study of N = 50 people, the disturbance variance equaled 2.0. I might then set α to (50)(.50) = 25 and β to (50)(.50)(2.0) = 50, which I have done in Line 9. You can increase the uncertainty in the prior distribution by using a smaller value for the previous study sample size when calculating the values of α and β and reduce the uncertainty by increasing the N. The following R code plots the inverse gamma distribution to give you a sense of its shape (note: you must install the R package `bayesAB` for this code to work):

```
library(bayesAB)
shape <- 25
scale <- 50
plotInvGamma(shape, scale)
```

Here is the resulting plot:



In Mplus, you can use mixtures of informative and uninformative prior distributions in the sense that Mplus will impose default uninformative prior distributions for all estimated parameters in the model but you then override selected defaults using the `MODEL PRIORS` command. For some models, you will need to posit multivariate prior distributions to ensure that the MCMC draws in the yield a positive definite covariance matrix. The inverse Wishart distribution is a common prior distribution in such cases. For additional perspectives on specifying informative prior distributions, see Asparouhov and Muthén (2010b), Gelman, Carlin and Stern (2013), van de Schoot, Sijbrandij, Depaoli, Winter, Olff, & van Loey (2018), and Zondervan-Zwijnenburg, Peeters, Depaoli, & van de Schoot (2017). For the case of small sample sizes, see Van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, and Van Loey (2015) and Smid et al. (2019). As a general rule, when working with informative prior distributions it is good practice to explore a range of different informative possibilities to determine how sensitive your results are to the prior values you chose.

Reducing Model Complexity for Small Sample Analysis

Another approach to making SEM amenable to small sample analysis is to simplify the model to the point that small sample statistics can be applied but without sacrificing the substantive or theoretical significance of the overall model. LISEM is one such approach where the model remains fully intact but estimation of parameter values and tests of model fit are pursued separately in different model segments. I discussed the strengths and weaknesses of this strategy in Chapter 8. It has both pluses and minuses.

A different approach to model simplification is to target the measurement model by turning multiple indicator latent variables into single indicator latent variables. With the

complexities of the measurement model eliminated, one might be able to apply FISEM or some variant of LISEM. This approach can be problematic in that it discards potentially useful information from the model but if the alternative is to not analyze the data at all, the compromise might be necessary. I discuss in Chapters 3 and 8 strategies to help mitigate the loss of information when using this approach.

A third strategy for reducing model complexity so that smaller sample sizes can be used is to eliminate variables from the model that are of lessor import or priority. I discussed strategies for mediator elimination in RETs in Chapter 17. Data reduction through factor analysis or principal components often is suggested as a way of reducing model complexity, but if all the measured variables are retained in the covariance matrix in order to derive factor or component loadings, then data reduction has not occurred, at least from the perspective of being amenable to small sample analysis; for elaboration, see Chapter 3.

For per equation analyses with single-indicator measures and a binary outcome, the classic methods of analysis in an LISEM framework are logit or probit regression. These methods rely on maximum likelihood, are based in asymptotic theory, and can be sample size demanding. The methods also can break down when the outcome is a rare event. A method known as **Firth regression** or **penalized maximum likelihood regression** (PMLE) has been suggested for logit and probit analyses with small samples or when the outcomes constitute rare events. An advantage of the PMLE method over another small sample approach, known as **exact logistic regression**, is that it is not as computationally intense, which sometimes mitigates against the use of the exact methods. PMLE deals with the small sample bias of traditional maximum likelihood by introducing (penalty) corrections to cancel the bias. For a primer on the method, see Cole, Chu and Greenland (2014). I provide a program for PMLE called *Small N logit/probit* on my website.

For continuous outcomes, OLS regression is the classic workhorse for many researchers in small sample LISEM scenarios, but a range of robust regression methods are viable competitors that better take into account outliers, do not make assumptions of normality or variance homogeneity, and that often have greater statistical power than OLS regression (Wilcox, 2021). I offer a program for one such method, MM regression, on my webpage but a host of other methods are described in Wilcox (2021), including robust variants for multi-level data, growth curves, analysis of covariance, robust logistic regression, and a range of other statistical approaches.

Small sample multi-level modeling methods outside of an SEM context have been described by McNeish (2017, 2017b) and McNeish and Stapleton (2016). I discuss them in Chapter 25. Small sample factor analysis methods have been reviewed by McNeish (2017c).

In sum, SEM is often characterized as a large sample statistical technique, but the method can be used for small sample sizes. To be sure, one must make some sacrifices as one deals with small sample sizes, but a small sample size does not necessarily rule out the use of SEM or variants of it.

CONCLUDING COMMENTS

Many researchers associate sample size decisions with the concept of statistical power, but sample size decisions are more complex than this. Not only must we think about practical and logistical constraints when making sample size decisions, we also must consider statistical power, robustness, margins of errors, asymptotic theory, covariance matrix stability, model complexity, effects on estimation properties, missing data, non-normality, and effect size sensitivity. Canned software for power analysis is limited in that it fails to address most of these issues; it keeps your focus narrow by considering only statistical power per se. By contrast, localized simulations, which are relatively easy to implement in Mplus or R, have the potential to provide information on all the above facets of sample size decision making.

Statistical power is impacted by sampling error, the strength of the targeted effect in the population, and your tolerance for Type I errors. Sample size only affects sampling error and sampling error is impacted by factors other than sample size (e.g., the variability of scores in the population). These facts open the door to more ways to increase power than just by increasing sample size. As researchers shift emphases away from statistical significance and more towards effect size estimation and effect size uncertainty/sensitivity, statistical power is becoming but one piece of the puzzle that one considers when making sample size decisions. I prefer to think about sample size in terms of its impact on effect size sensitivity and margins of error rather than my ability to reject an uninteresting null hypothesis that an effect is exactly zero. To be sure, Type I and Type II errors enter my thinking but more in the spirit of wanting to avoid erroneously concluding a *meaningful* effect exists when, in fact, it does not, or my missing the presence of a *meaningful* effect. In this sense, I focus my energies on defining what constitutes the minimum value for a meaningful effect for the variables I target and then I ensure I have sufficient effect size sensitivity to detect it with a reasonably low margin of error. To me, a strict emphasis on classic null hypothesis testing in the context of power analysis for making sample size decisions is an outdated approach to good program evaluation.

Localized simulations are a state of the art tool for assisting sample size decision making. They provide far more information than canned power analysis software and, importantly, they force you to think more comprehensively about the parameters in your

model. I use the approach by exploring variations in parameterizations and the assumptions I make in the simulations, which gives me a fuller picture of the implications of the parameterizations I choose and the assumptions I make. Although it takes some effort to master the design and execution of such simulations, I encourage you to do so as they represent a powerful tool for helping you design program evaluations.

APPENDIX: STANDARDIZED METRIC POPULATION VALUES

In this Appendix I describe how to choose population values for a simulation study using the standardized metric approach. I assume you have read the section in the main text on choosing population values for simulations for variables with raw metrics. I focus here on the model in Figure 28.1. An advantage of using standardized metrics (mean of zero and SD of 1) as raw scores is that they often make raw metric statistics more relatable while their patterns can produce the same power results as different metrics. For example, the statistical power for a two group (using, say, $n_1 = 65$, $n_2 = 65$) mean difference of 5.0 on an outcome that ranges from 0 to 100 with a pooled SD of 10.0 is the same as that for a two group mean difference of 0.50 with a pooled SD of 1.0. In both cases, the mean difference equals half a pooled standard deviation, hence the two cases produce identical statistical power. However, some researchers relate more to the latter case because the mean difference is analogous to a Cohen's *d* and because the power results for the standardized metric will generalize to any metric scenario where the mean difference is half the size of the pooled SD coupled with the other assumptions of the standardized metric analysis.

For the treatment dummy variable, T, scored 1 = the intervention group and 0 = the control group, the population variance and standard deviation is the same as in the main text for the standardized metric because the metric of T is unstandardized in all cases. For an equal number of participants per group, the population standard deviation of T always equals 0.50 and the variance is the square of this value, 0.25.

Because both study skills (which I will hereafter call M because it is the mediator) and exam performance (which I will hereafter call Y) are continuous variables, I arbitrarily set their standard deviations and variances to 1.0, per a standardized metric.

I next must decide how much of the variance in M that I want T to account for. This dictates both the value of the disturbance variance, d1, and the value of the path coefficient for T→M, or p_1. Suppose I want the effect size or value of p_1 to map roughly onto a medium effect size in Cohen's framework. This would be about 5% of explained variance or an eta squared of 0.05. This means that the disturbance variance for M should equal 1 (the variance of M) minus 0.05 = 0.95. Using the equation

$$var(M) = p_1^2\ var(T) + var\ (d_1)$$

I obtain

$$1.0 = p_1^2\ 0.25 + 0.95$$

and with simple algebraic manipulation of these numbers I find that

$p_1^2 = (1-.95) / 0.25 = 0.200$

and $p_1 = 0.447$

This exercise coupled with the previous steps yields the following population parameter values for the model thus far:

var(T)=0.25, var(M)=1, var(Y)=1, var(d_1)=0.95, and p_1=0.447

Next, I specify how much of the variance in Y that I want M to account for. Suppose I decide that the minimum effect size of M that I want to be sure to detect is one that reflects five percent explained variance in Y. If the proportion of explained variance in Y by M is 0.05, this means the disturbance variance for Y must equal 1 – 0.05 = 0.95. Using the equation

$var(Y) = p_2^2\ var(M) + var\ (d_2)$

I obtain

$1.0 = p_2^2\ 1.00 + 0.95$

and with simple algebraic manipulation I find

$p_2^2 = (1-.95) /1.00 = 0.05$

and $p_2 = 0.223$.

This gives me everything I need to conduct the simulation in terms of population parameters:

var(T)=0.25, var(M)=1, var(Y)=1, var(d_1)=0.95, var(d_2) = 0.95, p_1=0.447, p_2 = 0.223

I will evaluate power for a total sample size of 150.

Table A.1 presents the Mplus syntax to conduct the simulation. You should be familiar with all of it.

Table A.1: Local Simulation with Standardized Metric

```
1. TITLE: LOCAL SIMULATION ;
2. MONTECARLO:
3. NAMES ARE t m y ;
4. CUTPOINTS = t(0);
5. NOBS = 150 ;          !sample size
6. NREPS = 20000 ;       !number of replicates
```

```
7. SEED = 2222 ;        !random seed
8. !SAVE = temp.dat;
9. ANALYSIS:
10. ESTIMATOR = MLR ;
11. MODEL POPULATION:  !specify population model
12. [t*0] ;            !set mean of treatment to 0 for cutoff
13. t*0.25 ;           !define var of treatment variable
14. [y*0]; [m*0];      !set intercepts to 0
15. y ON m*.223 ;      !set effect of m on y
16. m ON t*.447 ;      !set effect of t on M
17. y*.95 ;            !disturbance variance for y
18. m*.95 ;            !disturbance variance for m
19. MODEL:             !specify analysis model
20. y ON m*.223 ;      !outcome equation
21. m ON t*.447 ;      !mediation equation
22. y*.95 ;            !disturbance variance for y
23. m*.95 ;            !disturbance variance for m
24. MODEL INDIRECT:
25. y IND t ;          !evaluate omnibus mediation effect
26. OUTPUT: TECH9 ;
```

The output follows the same formatting as that for the raw metric discussed in the main text.