Significance Testing Primer

This primer addresses statistical significance testing and the technique of bootstrapping, with special attention to their roles in SEM. Significance testing has become increasingly controversial over the years. This is due both to the inherent limitations of significance testing and to the failure of most researchers to understand what statistical significance means. Estimation of confidence intervals (interval estimation) as an alternative to significance testing is described. Two different methods for calculating confidence intervals for statistics with complex distributions are outlined: noncentrality interval estimation and bootstrapping. Some types of fit statistics in SEM are distributed as noncentral test statistics, and bootstrapping is a computer-based resampling procedure with application in SEM.

STANDARD ERRORS

The standard error is a standard deviation in a **sampling distribution**, the probability distribution for a sample statistic based on all possible random samples selected from the same population and each based on the same N. The standard error estimates **sampling error**, or the difference between sample statistics and the corresponding population parameter. Given constant variability among cases, standard error varies inversely with N. This means that distributions of statistics from larger samples are generally narrower (less variable) than distributions of the same statistic from smaller samples.

There are textbook formulas for standard errors of statistics with simple distributions. By "simple" I mean that (1) the statistic estimates a single parameter and (2) the basic shape of its distribution does not change as a function of that parameter. For example, means have simple distributions, and the equation for their standard error is

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \tag{S.1}$$

where σ is the population standard deviation among cases. Given σ, the value of σ_M decreases as N increases (see Figure S.1). An original normal distribution along with two different sampling distributions of means for $N = 5$ and $N = 25$ are depicted. There is greater variation of sample means around the population mean μ when the sample size is smaller. The value of σ_M must be estimated, if σ is unknown. The estimator is

$$SE_M = \frac{SD}{\sqrt{N}} \tag{S.2}$$

Note that SE_M itself has a standard error. This is because the value of SE_M will vary over random samples drawn from the same population.

Standard errors for statistics from observed variables estimate sampling error under the exacting assumptions stated next:

1. The method of sampling is random, or at least haphazard enough to generate representative samples over replications.
2. There is no other source of error besides sampling error.

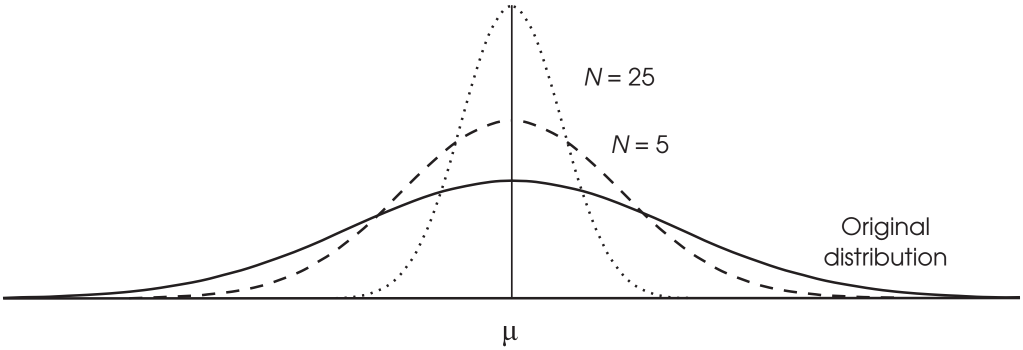

FIGURE S.1. An original distribution of scores and two distributions of random sample means each based on different sample sizes, N = 5 and N = 25.

3. Standard errors for parametric statistics often assume normality or homoscedasticity.

The problem with the assumptions just stated is that they are false in most studies. For example, true random sampling requires a list of all members in a population, but such lists are rare. Most samples in human research are ad hoc (convenience) samples made up of participants who happen to be available. What standard errors measure in such samples is generally unknown. Scores are affected by multiple types of error, including sampling error, measurement error, and, in treatment outcome studies, implementation error, or deviations from a treatment protocol, such as due to poor patient or therapist compliance. Other types of error include specification error, such as left-out variables error, and a host of threats to internal validity (e.g., confounding), external validity (e.g., interference due to multiple treatments), and construct validity (e.g., scores are not reliable) (Shadish, Cook, & Campbell, 2001). But standard errors generally assume that the scores are perfect in every way except for the vagaries of random sampling.

The normality assumption refers to population distributions, but normal distributions in actual studies are rare. Many, if not most, empirical distributions are not even symmetrical, much less normal, and departures from normality are often strikingly large (Micceri, 1989). Geary (1947, p. 214) suggested that the disclaimer, "Normality is a myth; there never was, and never will be, a normal distribution," should appear in all statistics textbooks. Ratios across different groups of largest to smallest variances as large as 8:1 are not uncommon (Keselman et al., 1998), so perhaps homoscedasticity is a myth, too. Even small departures from distributional assumptions can appreciably distort standard errors in small or unrepresentative samples. There are robust estimators with fewer distributional assumptions (Erceg-Hurn & Mirosevich, 2008), but their standard errors assume random sampling, too.

CRITICAL RATIOS

The basic form of a significance test is the **critical ratio**, the ratio of a statistic over its standard error. Assuming large samples and normality, a critical ratio is interpreted as a deviate in a normal curve (z) with a mean of zero and a standard deviation that equals the standard error. A heuristic is that if $|z| > 2.00$, the null hypothesis (H_0) that the corresponding parameter is zero is rejected at the .05 level ($p < .05$) for a two-tailed test (H_1). The precise value of $|z|$ for the .05 level is 1.96, and for the .01 level it is 2.58. For example, given

$$M = 5.00,\ SD = 25.00,\ N = 100,$$
$$H_0\text{: } \mu = 0, \text{ and } H_1\text{: } \mu \neq 0$$

$$SE_M = \frac{25.00}{\sqrt{100}} = 2.50 \quad \text{and} \quad = \frac{5.00}{2.50} = 2.00$$

For z = 2.00 and assuming random sampling and no other error besides sampling error, p = .046, so the null hypothesis that the population mean is zero is rejected at the .05 level.

In small samples, the ratio M/SE_M approximates a t distribution, which necessitates the use of special tables to determine the critical values of t for the .05 or .01 levels.[1] These distributions are **central *t* distributions** where the null hypothesis is assumed to be true. Such distributions have a single parameter, *df*, the degrees of freedom, which are $N - 1$ for a single mean. Other forms of the t test for means have different *df* values. For instance, $df = N - 2$, where N is the total number of cases when means from two independent samples are compared, but all central t distributions assume a true null hypothesis. There are central distributions for other test statistics, such as F or χ^2, and tables of critical values for these familiar test statistics can be found in many statistics textbooks and also online.

In some SEM computer programs, standard errors are calculated for the unstandardized solution only. You can see this fact when you look through the computer output and find no standard errors printed for standardized estimates. This means that results of significance tests (z) are available only for the unstandardized estimates. Researchers often assume that p values for unstandardized estimates also apply to the corresponding standardized estimates. For samples that are large and representative, this assumption may not be unreasonable. But you should know that the p value for an unstandardized estimate does not automatically apply to its standardized counterpart. This is because standardized estimates have their own standard errors, and their critical ratios may not correspond to the same probabilities as the critical ratios for the corresponding unstandardized results. This explains why you should (1) always report the unstandardized solution including the standard errors and (2) not associate p values for unstandardized estimates with the corresponding standardized estimates. An example follows.

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

✗ .60*

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

✓ 4.20* (2.10) .60

where the asterisk is associated with the unstandardized estimate (4.20), not the standardized one (.60).

POWER AND TYPES OF NULL HYPOTHESES

The failure to reject a null hypothesis, such as $p \geq .05$ when testing at the .05 level, is meaningful only if (1) the power of the test is adequate and (2) the null hypothesis is at least plausible to some degree. **Power** is the probability of getting statistically significant results over random samples when the null hypothesis is false. Power is also the complement of the probability of a Type II error (failing to reject H_0 when it is false), often designated as β, so $1 - \beta$ = power. Whatever increases power decreases β, and vice versa. Power varies directly with the magnitude of the population effect size and your sample size. Other factors that affect power include:

1. The level of statistical significance α (e.g., .05 vs. .01) and the directionality of H_1 (i.e., one- or two-tailed tests).
2. Whether samples are independent or dependent (i.e., between-subjects or within-subjects design).
3. The particular test statistic used.
4. The reliability of the scores.

The following combination generally leads to the greatest power: a large sample, specification of $\alpha = .05$, a one-tailed (directional) H_1, a within-subjects design, a parametric test statistic (e.g., t) rather than a nonparametric statistic (e.g., Mann–Whitney U), and scores that are very reliable.

Power should be estimated when the study is planned but before the data are collected. Some granting agen-

[1] Within large samples, t and z for the same statistic are essentially equal, and their values are asymptotic in very large samples.

cies require such a priori estimates of power in applications for funds. If power is low, there is little point in carrying out the study, if outcomes of significance testing are important. For example, if power is only about .50, then the likelihood over random samples of rejecting a false null hypothesis is no greater than guessing the outcome of a coin toss. In this case, tossing a coin instead of conducting the study would be just as likely to give the correct decision in the long run and would save time and money, too.

Unfortunately, only about 10% of researchers report the a priori power of their analyses (Ellis, 2010). This is a problem because without knowing power estimates, one is unable to correctly interpret results that are not statistically significant. That is, do such results indicate lack of support for the researcher's hypothesis or just the expected consequence of inadequate power? There is free software for power analysis, so the widespread failure to estimate and report power is bewildering.[2] How to estimate power in SEM is described in Chapter 10, but power for certain kinds of significance tests in SEM is often quite low even in large samples.

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

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

which predicts that two population means are exactly equal. The problem with nil hypotheses is that it is unlikely that the value of any parameter (or difference between two parameters) is exactly zero, especially if zero means the total absence of an effect or association. It is possible for the *t* test to specify a **non-nil hypothesis**, such as

$$H_0\colon \mu_1 - \mu_2 = 5.0$$

but doing so is rare in practice. As the name suggests, a non-nil hypothesis predicts that a population effect or association is not zero.

It is more difficult to specify and test non-nil hypotheses for other test statistics, such as *F* when comparing three or more means. This is because computer programs almost always assume a nil hypothesis. Nil hypotheses may be appropriate in new research areas where it is unknown whether effects exist at all, but such hypotheses are less suitable in more established areas where it is already known that certain effects are not zero. If so, then (1) an implausible nil hypothesis is an uninteresting "straw man" argument (a fallacy) that is easily rejected, and (2) *p* values in significance testing are too low. This happens because the data seem more exceptional than they really are compared with evaluating the same data under a more realistic non-nil hypothesis.

SIGNIFICANCE TESTING CONTROVERSY

Until recently, significance testing was both routine and expected (i.e., everybody did it). But significance testing has been increasingly criticized as unscientific and unempirical (Kline, 2013; Lambdin, 2012). Some authors in **statistics reform** suggest that overreliance on significance testing can lead to **trained incapacity**, or the inability of researchers to understand their own results due to inherent limitations of significance tests and myriad associated cognitive distortions (Ziliak & McCloskey, 2008). Essential criticisms of significance testing are listed next:

1. Outcomes of significance tests—*p* values—are wrong in most studies.
2. Researchers do not understand *p* values.
3. Most applications of significance testing are incorrect.
4. Significance tests do not tell researchers what they want to know.

The fact that *p* values are calculated under implausible assumptions (e.g., random sampling, normality, no measurement error) was mentioned earlier in the section on standard errors. Distributional assumptions are rarely verified because researchers mistakenly believe that significance tests are robust even in small, unrepresentative samples (Hoekstra, Kiers, & Johnson, 2013). If assumptions are checked, the wrong methods are used, including significance tests that supposedly verify distributional assumptions of other significance tests, such as Levene's test for homoscedasticity. The problem with such tests is that their results are often wrong due in part to their own unrealistic assumptions (Erceg-Hurn & Mirosevich, 2008).

Most researchers misinterpret statistical significance.

[2] *https://www.gpower.hhu.de/en.html*

For example, about 80–90% of psychology professors endorse false beliefs about statistical significance, no better than psychology undergraduate students in introductory statistics courses (Haller & Krauss, 2002). These comparably high rates of misinterpretation suggest an ongoing cycle of misinformation, where instructors or text books transmit false information to students, who then perpetuate the myths to the next generation. Most false beliefs about p values involve overinterpretation that favor the researcher's hypotheses, which is a form of confirmation bias—see Topic Box S.1 for a review of the "Big Five" misinterpretations of statistical significance. Exercises 1–3 ask you to comment on examples of incorrect definitions of p values.

Most researchers fail to report the power of their significance tests. Another misuse comes from treating the conventional levels of statistical significance, .05 or .01, as golden rules that apply to all studies and disciplines. The value of α sets the risk of Type I error, or the probability over random samples that a true null hypothesis

TOPIC BOX S.1

Cognitive Errors in Significance Testing

First, we consider the correct interpretation of p values, which is actually quite narrow in scope. They represent the conditional probability:

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

which is the likelihood of a sample result or one even more extreme assuming random sampling under a true null hypothesis and where all other assumptions are met (distributional requirements, no error other than sampling error, independent and perfectly reliable scores, etc.). Most of what contributes to a p value are those *even more extreme* results that were not actually observed; that is, p values are only partially empirical. Two correct interpretations for the case $p < .05$ are given next. Other correct definitions are probably just variations of the ones that follow:

1. Suppose the study were repeated by drawing many random samples from the same population(s) where the null hypothesis is true (i.e., every result happens by chance). Less than 5% of these hypothetical results would be even more inconsistent with H_0 than the actual result.
2. Less than 5% of test statistics from many random samples are further away from the mean of the sampling distribution under H_0 than the one for the observed result. In other words, the odds are less than 1 to 19 of getting a result from a random sample even more extreme than the observed one.

Described next are what I call the "Big Five" misinterpretations of p values. The **odds against chance fallacy** is the false belief that p indicates the probability that a particular result happened by chance (i.e., due to sampling error). Remember that p is calculated for a range of results, most unobserved, and not for any single result. Also, p is calculated assuming that H_0 is already true, so the probability that sampling error is the only explanation is already taken to be 1.0. Thus, it is illogical to view p as measuring the likelihood of sampling error. Besides, the probability that sample results are affected by error of some kind—sampling, measurement, or specification error, among others—is virtually 1.0. From this perspective, virtually all sample results are wrong (Ioannidis, 2005). That is, our data routinely lie, they lie through multiple types of error, and it is only when results are averaged over studies, such as in the technique of

(continued)

meta-analysis, that some of these errors begin to cancel out. Significance testing in individual studies in no way helps in this process.

The **local Type I error fallacy** for the case where $p < .05$ and $\alpha = .05$ (i.e., H_0 is rejected) says that the likelihood that the decision just taken to reject the null hypothesis is a Type I error is less than 5%. This belief is false because any particular decision to reject H_0 is either correct or incorrect, so no probability (error other than 0 or 1.0) is associated with it. Only with sufficient replication could we determine whether or not the decision to reject H_0 in a particular study was correct. The **inverse probability fallacy** is the false belief that *p* is the probability that the null hypothesis is true. This error stems from forgetting that *p* values are probabilities of data under the null hypothesis, not the other way around.

Two other fallacies concern the complements of *p* values, or $1 - p$. The **valid research hypothesis fallacy** is the false belief that $1 - p$ is the probability that the alternative hypothesis is true. The quantity $1 - p$ is a probability, but it is just the likelihood of getting a result even *less* extreme under H_0 than the one actually found. The **replicability fallacy** is that $1 - p$ is the likelihood of finding the same result in another random sample. If this fallacy were true, knowing the likelihood of replication would be very useful. Unfortunately, *p* is just the probability of data in a particular sample under a specific null hypothesis. In general, replication is a matter of experimental design and whether some effect actually exists in the population (i.e., it is an empirical question). Kline (2013, chap. 4) describes additional false beliefs about *p* values.

will be rejected, but Type II error is often more serious. An example is when a nil hypothesis is known to be false before even collecting the data. In this case, the effective level of α is zero, and Type II error is the only possible kind of error. Another example is when a treatment for an illness is beneficial, but the results are not significant at $p < .05$, the highest conventional level of α. Type II error in this context means that a beneficial treatment is not detected. There is actually no requirement to specify an arbitrary level of α (i.e., .05 or .01) that does not properly balance the risk of Type I error against that of Type II error (Hurlbert & Lombardi, 2009).

Armstrong (2007) argued that significance testing does not foster progress in science even if such tests are properly conducted. This is because their results do not tell researchers what they wish to know, including the likelihood that some hypothesis is true, given the data; the probability that a Type I error has occurred, given that the null hypothesis was just rejected; the prospects for replication; and whether the findings are actually important. An alternative is to describe replicated results in terms of their effect sizes and precisions (confidence intervals) and interpret their substantive significance using language relevant to stakeholders in a particular research context (Aguinis et al., 2010). Given all the problems just considered, significance testing is actually banned in some research journals such as *Basic and Applied Social Psychology* (Trafimow & Marks, 2015). See also the special edition on significance testing in the journal *American Statistician* (Wasserstein et al., 2019).

CONFIDENCE INTERVALS AND NONCENTRAL TEST DISTRIBUTIONS

Interval estimation is an alternative to significance testing. It involves reporting effect sizes with confidence intervals (error bars, margins of error) that indicate a range of results considered equivalent within the limits of sampling error to the specific result found (i.e., the point estimate). For statistics with simple distributions, the width of either side of a

$$100 \times (1 - \alpha)\ \%$$

confidence interval is determined by the product of the standard error and the critical value of a central test statistic at the α level of statistical significance for a two-tailed alternative hypothesis. For example, given

$$M = 100.00,\ SD = 9.00,\ N = 25, \text{ and } SE_M = 1.80$$

the 95% confidence interval is

$$100.00 \pm (1.80)\ t_{\text{2-tail, } \alpha = .05}\ (24)$$

where $t_{\text{2-tail, } \alpha = .05}$ (24) is the positive two-tailed critical value in a central *t* distribution at the .05 level of statistical significance, which for $df = 24$ is 2.064.[3] The 95% confidence interval is thus

$$100.00 \pm 1.80\ (2.064), \text{ or } 100.00 \pm 3.72$$

which defines the interval [96.28, 103.72]. This interval specifies a range of values considered equivalent to the observed mean within the limits of sampling error at the 95% confidence level. The point estimate of 100.00 falls at the exact center of the interval, and the whole interval explicitly conveys the idea that a margin of error is associated with the corresponding statistic (100.00). Note that the interval [96.28, 103.72] is based on a single estimate of σ_M, or $SE_M = 1.80$. But this quantity (1.80) is itself just a point estimate, and the value of SE_M in a different sample will almost certainly not be 1.80. This means that the interval [96.28, 103.72] is actually too narrow (i.e., more precise than it seems), if we also consider sampling error in SE_M.

Because confidence intervals are based on the same standard errors as significance tests—and rely on the same unrealistic assumptions—researchers should not overinterpret their lower or upper bounds. Suppose a 95% confidence interval based on $M = 2.50$ is [0, 5.00], which includes zero. This fact can be misinterpreted, such as wrongly concluding that $\mu = 0$. But zero is only one value within a range of estimates, so it has no special status. This means that the hypothesis that $\mu = 0$ is not favored any more than the hypothesis that $\mu = 5.00$ (or that μ equals any other value in the range 0–5.0). Confidence intervals are subject to sampling error, too, so zero may not fall within the 95% confidence interval in a replication sample. Do not believe that confidence intervals are just significance tests in disguise (Thompson, 2006). This is because null hypotheses are required for significance tests, but not for confidence intervals, and many null hypotheses have little scientific value.

Statistics with complex distributions may not follow central distributions. For example, if $\rho^2 = 0$ (i.e., the squared population correlation is zero), then distributions of R^2 follow **central *F* distributions** with *k* and $N - k - 1$ degrees of freedom, where *k* is the number of predictors. Central *F* distributions assume $\rho^2 = 0$ and provide the critical values for the familiar *F* test in multiple regression or ANOVA. But if $\rho^2 > 0$, the sampling distribution for R^2 is defined by **noncentral *F* distributions**, which have an additional parameter, called the **noncentrality parameter.** This parameter indicates the degree to which the null hypothesis that $\rho^2 = 0$ is false. Noncentral *F* distributions take the form

$$F\ (k, N - k - 1, \lambda) \tag{S.3}$$

where λ is the noncentrality parameter. The latter is related to ρ^2 and the sample size, or

$$\lambda = N\left(\frac{\rho^2}{1-\rho^2}\right) \tag{S.4}$$

If $\rho^2 = 0$, then $\lambda = 0$, which indicates no departure from the nil hypothesis. Equation S.4 can be rearranged to express ρ^2 as a function of λ and the sample size:

$$\rho^2 = \frac{\lambda}{N + \lambda} \tag{S.5}$$

Presented in Figure S.2 are two *F* distributions where the degrees of freedom are 5 and 20. For the central *F* distribution in the left part of the figure, $\lambda = 0$. But $\lambda = 10.0$ for the noncentral *F* distribution in the right side of the figure. Note in the figure that (1) both distributions are positively skewed, but the central *F* distribution has greater skew than the noncentral *F* distribution. Also, (2) the noncentral *F* distribution has a greater expected value—the weighted average of all possible values—than the central *F* distribution. This is because the noncentral *F* distribution in the figure assumes that $\rho^2 > 0$, but the central *F* distribution is for $\rho^2 = 0$.

Steiger and Fouladi (1997) showed that if we can obtain a confidence interval for λ, we can also obtain a confidence interval for ρ^2 using Equation S.5. To do so, we use a computer tool that finds λ_L, the lower bound of the confidence interval for λ. For the 95% level, the lower bound λ_L equals the value of λ for the noncentral *F* distribution in which the observed *F* falls at the 97.5th percentile. The upper bound λ_U equals the value of λ for the noncentral *F* distribution in which the

[3] See the calculating webpage at *https://www.usablestats.com/calcs/tinv*

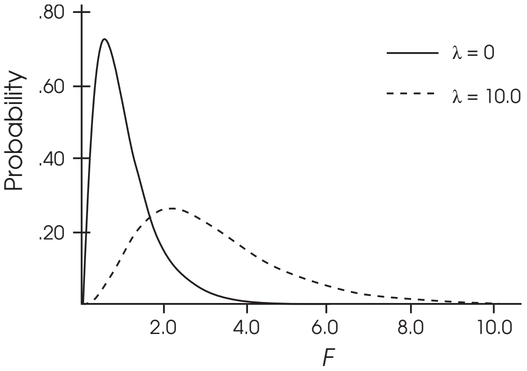

FIGURE S.2. Distributions of central *F* and noncentral *F* for 5 and 20 degrees and where the noncentrality parameter (λ) equals 0 for central *F* and λ = 10.0 for noncentral *F*.

observed *F* falls at the 2.5th percentile. But we need to find which particular noncentral *F* distributions are the most consistent with the data, and this is the problem solved with the right computer tool. An example follows.

I used J. Steiger's Noncentral Distributional Calculator (NDC), a freely available Windows application for noncentrality interval estimation.[4] For the data in Table R.1 (see the Regression Primer)

$$R^2_{Y \cdot X,\, W} = .576,\ N = 20, \text{ and } F\,(2,\, 17) = 11.536$$

We can say the observed *F* of 11.536 falls at the

1. 97.5th percentile in the noncentral *F* (2, 17, 4.190) distribution; and the same observed *F* falls at the
2. 2.5th percentile in the noncentral *F* (2, 17, 52.047) distribution.

For these data, the 95% confidence interval for λ is [4.190, 52.047]. Using Equation S.5 to convert the lower and upper bounds of this interval to ρ^2 units for $N = 20$ gives us the noncentral 95% confidence interval based on $R^2 = .576$, which is [.173, .722]. (You should verify these results.) The interval just reported is not symmetrical about $R^2 = .576$, but this is expected in noncentrality interval estimation. Exercise 4 asks you to calculate the 95% noncentral confidence interval based on the same value of R^2 but in a larger sample.

There are noncentral distributions for other test statistics, such as *t* and χ^2, and they all assume that the null hypothesis is false by the degree indicated by the value of the noncentrality parameter. The latter equals zero in central test distributions, so central test distributions are just special cases of noncentral test distributions (i.e., they belong to the same distribution family). Noncentral test distributions play an important role in certain types of statistical analyses. Computer programs that estimate the power of significance tests as a function of study characteristics and the predicted effect size analyze noncentral distributions. This is because the concept of power assumes that the null hypothesis is false, and it is false by the degree indicated by a nonzero effect size. The latter generally corresponds to a value of the noncentrality parameter that is also not zero.

Another application is the estimation of confidence intervals based on sample statistics that measure effect size besides R^2. For example, distributions of standardized mean differences (*d*), or the ratio of a mean contrast over the standard deviation, generally follow central *t* distributions when the corresponding parameter is zero; otherwise, *d* statistics are distributed as noncentral *t* distributions. There are special computer programs for noncentrality interval estimation based on *d* statistics (Cumming & Calin-Jageman, 2017). Effect size estimation also generally assumes that the null hypothesis—especially when it is a nil hypothesis—is false.

[4] *http://www.statpower.net/Software.html*

Some measures of model fit in SEM are based on noncentral χ^2 distributions. These statistics measure the degree of **approximate (close) fit**, which allow for an "acceptable" amount of departure from **exact (perfect) fit**. What is considered "acceptable" departure from perfection is related to the value of the noncentrality parameter for the χ^2 that the computer calculates for the model and data. Other fit statistics in SEM measure the departure from exact fit, and these statistics are generally described by central χ^2 distributions, where the null hypothesis that the model has perfect fit in the population is assumed to be true. But the null hypothesis just stated is assumed to be false by statistics that measure approximate fit. Assessment of model fit against these two standards, approximate versus exact, is covered later in Chapter 10.

BOOTSTRAPPING

The technique of bootstrapping was developed by the statistician B. Efron in the 1970s (e.g., 1979). It is a computer-based method of **resampling** that combines the cases in a data set in different ways to estimate statistical precision. Perhaps the best known form is **nonparametric bootstrapping**, which generally makes no assumptions other than that the distribution in the sample reflects the basic shape of that in the population. This method treats your sample (i.e., data file) as a pseudo-population in that cases are randomly selected *with replacement* to generate other data sets, usually of the same size as the original. Because of sampling with replacement, (1) the same case can be selected in more than one generated data set or at least twice in the same generated sample, and (2) the composition of cases will vary slightly across the generated samples.

When repeated many times (e.g., 500) by the computer, bootstrapping simulates random sampling with replacement. It also constructs an **empirical sampling distribution**, the frequency distribution of the values of a statistic across generated samples. **Nonparametric bootstrapped confidence intervals** are calculated in the empirical distribution. For example, the lower and upper bounds of a 95% bootstrapped confidence interval correspond to, respectively, the 2.5th and 97.5th percentiles in the empirical sampling distribution. These limits contain 95% of the bootstrapped values of the statistic. This method is potentially useful for statistics with complex distributions. An example follows.

I used the nonparametric Bootstrap procedure of SimStat for Windows (Version 2.6.1) (Provalis Research, 1995–2011) to resample from the data in Table R.1 (see the Regression Primer) in order to generate a total of 500 bootstrapped samples each with 20 cases.[5] Presented in Figure S.3 is the empirical sampling distribution of R^2 across all generated samples. SimStat reported that the mean of this distribution is .626, the median is .630, and the standard deviation is .102. The first result (.626) is close to the observed value of R^2 = .576 for these data, which is expected.

The nonparametric bootstrapped 95% confidence interval in the empirical sampling distribution for this

[5] *https://provalisresearch.com*

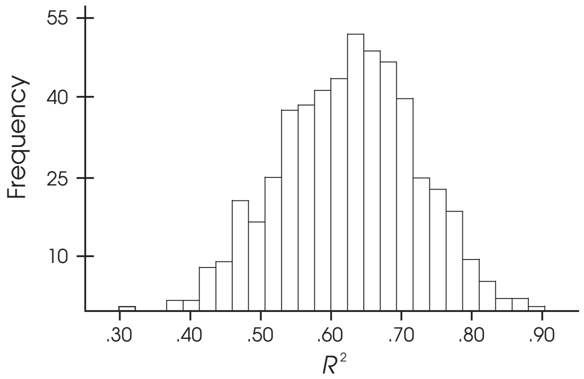

FIGURE S.3. Empirical sampling distribution for $R^2_{Y \cdot X,W}$ in 500 bootstrapped samples for the data in Table R.1.

example is [.425, .813]. This result from bootstrapping is quite different from the noncentral 95% confidence interval we calculated earlier for the same data, or [.173, .722], but bootstrapped results in small samples can be very inaccurate. This is because bootstrapping can magnify the effects of unusual features in a small data set. Note that the computer will generate a different empirical sampling distribution for the same data, if each time it is given a different **seed**, or a long number (vector) used to initiate simulated random sampling. Consequently, any result in a single application of nonparametric bootstrapping is not generally unique.

A raw data file is needed for nonparametric bootstrapping. This is not true in **parametric bootstrapping**, where the computer randomly samples from a theoretical probability density function specified by the researcher. When repeated many times by the computer, values of statistics in synthesized samples vary randomly about the specified parameters, which simulates sampling error. Parametric bootstrapping is a kind of Monte Carlo method that is used in computer simulation studies of the properties of estimators. Distributional assumptions can be added incrementally in parametric bootstrapping or successively relaxed over the generation of synthetic data sets.

Several SEM computer tools, including Amos, EQS, LISREL, Mplus, Stata, and `lavaan` for R, feature bootstrap methods. Some of these methods can estimate standard errors or generate confidence intervals based on certain estimators, such as statistics that measure model–data correspondence or indirect causal effects (Hancock & Liu, 2012). Parametric bootstrapping methods are used in SEM to conduct simulation studies, such as for power analysis, sample size determination, and hypothesis testing (Bandalos & Gagné, 2012).

SUMMARY

Statistical significance is not a gold scientific standard, and thinking about data analysis as a search for whether results are "significant" or "not significant" may be fruitless. This is because the presence of statistical significance does not reliably signal that results are noteworthy or even of mild interest, just as the failure to find statistical significance does not indicate that nothing of interest was found. It is also true that many, and perhaps most, researchers do not understand what statistical significance really means. Researchers should instead think more about whether observed effect sizes are precise and large enough to be of substantive interest. Keeping a skeptical view of significance testing will help you in SEM—and in other kinds of complex multivariate analyses, too—to avoid getting lost in a blizzard of asterisks. Also reviewed in this primer was the logic of noncentrality interval estimation and bootstrapping, both of which can be used to calculate confidence intervals based on statistics with complex distributions, including some that are used in SEM.

LEARN MORE

Kline (2013, chap. 4) describes additional cognitive errors about statistical significance, and Lambdin (2012) and Ziliak and McCloskey (2008) offer strong critiques of significance testing.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences.* Washington, DC: American Psychological Association.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory and Psychology, 22*, 67–90.

Ziliak, S., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives.* Ann Arbor: University of Michigan Press.

EXERCISES

Explain what is wrong versus right in each definition of statistical significance listed next.

1. The statistical significance of a result is an estimated measure of the degree to which it is true (in the sense of "representative of the population"). More technically, the value of the p level represents a decreasing index of the reliability of a result. The higher the p level, the less we can believe that the observed relation between variables in the sample

is a reliable indicator of the relation between the respective variables in the population. Specifically, the *p*-level represents the probability of error that is involved in accepting our observed result as valid, that is, as "representative of the population."

2. This is a very important and common term in psychology, but one that many people have problems with. Technically, statistical significance is the probability of some result from a statistical test occurring by chance. . . . Most often, psychologists look for a probability of 5% or less that the results are due to chance, which means a 95% chance the results are "not" due to chance.[6]

3. The calculation of statistical significance is subject to a certain degree of error. The researcher must define in advance the probability of a sampling error. Sample size is an important component of statistical significance in that larger samples are less prone to flukes. Only random, representative samples should be used in significance testing.

4. Calculate the 95% noncentral confidence interval for $R^2_{Y \cdot X,\,W}$ = .576, F (2, 47) = 31.925, and N = 50 using a computer tool for noncentrality interval estimation.

[6] *https://www.alleydog.com/glossary/definition.php?term=Statistical+Significance*

ANSWERS

Comments about the selected quotes:

1. Representativeness is determined by how cases are selected, which has nothing to do with statistical significance. If "reliability" means "repeatability," then statistical significance does not directly indicate the likelihood of replication. But if "reliability" means "sampling error," then, yes, there is less sampling error over larger random samples. Also, *p* is not the probability of error, which is virtually 1.0 for sample results, and neither is p the probability that the null hypothesis is true.

2. This is a restatement of the odds against chance fallacy. A *p* value does not indicate the likelihood that a particular result is due to chance, nor does 1 – *p* measure the probability that the data are due to any "real" effect. All sample results are affected by error.

3. The probability of sampling error is virtually 1.0 and thus cannot be specified in advance. The level of α is specified by the researcher in advance, but there is actually no requirement to specify an arbitrary criterion level of statistical significance. The rest of the quote is correct, including the claim that significance testing assumes random sampling.

4. I used the NDC calculator for this problem. We can say that F (2, 47) = 31.925 falls at

   a. 97.5th percentile in the noncentral F (2, 47, 28.573) distribution; and the same observed F falls at the

   b. 2.5th percentile in the noncentral F (2, 17, 109.201) distribution.

   So the 95% confidence interval for λ is [28.573, 109.201]. Using Equation S.4 to convert the lower and upper bounds of this interval to ρ^2 units for N = 50 gives us the noncentral 95% confidence interval based on R^2 = .576, which is [.364, .686]. As expected, this interval is narrower than the corresponding interval based on N = 20, which is [.173, .722].

REFERENCES

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods, 13*(3), 515–539.

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting, 23*(2), 321–327.

Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 92–108). New York: Guilford Press.

Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 625–666). IAP.

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*(1), 1–26.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591–601.

Geary, R. C. (1947). Testing for normality. *Biometrika, 34*(3–4), 209–242.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1–17.

Hancock, G. R., & Liu, M. (2012). Bootstrapping standard errors and data–model fit statistics in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 277–295). Guilford Press.

Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2013). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology, 3*. Retrieved from *www.frontiersin.org/article/10.3389/fpsyg.2012.00137/full*

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review, 21*(5), 1157–1164.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman–Pearson decision theory framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46*(5), 311–349.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), Article 124.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of education researchers: An analysis of the ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*(3), 350–368.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). American Psychological Association.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology, 22*(1), 67–90.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166.

Provalis Research. (1995–2011). SimStat (Version 2.6.1) [Computer software]. https://provalisresearch.com/

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J, H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Erlbaum.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. Guilford Press.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1–2.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$." *American Statistician, 73*(Suppl. 1), 1–19.

Ziliak, S., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.