CHAPTER 1

THE NEW STATISTICS

1.1 REQUIRED BACKGROUND

This book begins with analyses that involve three variables, for example, an independent variable, a dependent variable, and a variable that is statistically controlled when examining the association between these, often called a covariate. Later chapters describe situations that involve multiple predictors, multiple outcomes, and/or multiple covariates. The bivariate analyses covered in introductory statistics books are the building blocks for these analyses. Therefore, you need a thorough understanding of bivariate analyses (i.e., analyses for one independent and one dependent variable) to understand the analyses introduced in this book.

The following topics are covered in Volume I (*Applied Statistics I: Basic Bivariate Techniques* [Warner, 2020]) and most other introductory statistics books. If you are unfamiliar with any of these topics, you should review them before you move forward.

- The use of frequency tables, histograms, boxplots, and other graphs of sample data to describe approximate distribution shape and extreme outliers. This is important for data screening.
- Understanding that some frequently used statistics, such as the sample mean, are not robust against the impact of outliers and violations of other assumptions.
- Computing and interpreting sample variance and standard deviation and the concept of degrees of freedom (*df*).
- Interpretation of standard scores (*z* scores) as unit-free information about the location of a single value relative to a distribution.
- The concept of sampling error, indexes of sampling error such as SE_M, and the way sampling error is used in setting up confidence intervals (CIs) and statistical significance tests.
- Choice of appropriate bivariate statistics on the basis of types of variables involved (categorical vs. quantitative and between-groups designs vs. repeated measures or paired or correlated samples).
- The most commonly used statistics, including independent-samples *t*, between-*S* analysis of variance (ANOVA), correlation, and bivariate regression. Ideally, you should also be familiar with paired-samples *t* and repeated-measures or paired-samples ANOVA. The multivariate and multivariate analyses covered in this book are built on these basic analyses.
- The logic of statistical significance tests (null-hypothesis statistical testing [NHST]), interpretation of *p* values, and limitations and problems with NHST and *p* values.

- Distributions used in familiar significance tests (normal, t, F, and χ^2) and the use of tail areas to describe outcomes as unusual or extreme.
- The concept of variance partitioning. In correlation and regression, r^2 is the proportion of variance in Y that can be predicted from X, and $(1 - r^2)$ is the proportion of variance in Y that cannot be predicted from X. In ANOVA, $SS_{between}$ provides information about proportion of variance in Y that is predictable from group membership, and SS_{within} provides information about variance in Y that is not predictable from group membership.
- Effect size.
- The difference between statistical significance and practical or clinical importance.
- Factors that influence statistical power, particularly effect size and sample size.

1.2 WHAT IS THE "NEW STATISTICS"?

In the past, many data analysts relied heavily on statistical significance tests to evaluate results and did not always report effect size. Even when used correctly, significance tests do not tell us everything we want to know; misuse and misinterpretation are common (Greenland et al., 2016). Misuse of significance tests has led to selective publication of only results with $p < .05$; publication of these selected results has sometimes led to widespread reports of "findings" that are not reproduced when replication studies are performed. The focus on "new" and "statistically significant" outcomes means that we sometimes don't discard incorrect results. Progress in science requires that we weed out mistakes, as well as make new discoveries.

Proponents of the "New Statistics" (such as Cumming, 2014) do not claim that their recommendations are really new. Many statisticians have called for changes in the way results are evaluated and reported, at least since the 1960s (including but not limited to Cohen, 1988, 1992, 1994; Daniel, 1998; Morrison & Henkel,1970; and Rozeboom, 1960). However, practitioners of statistics are often slow to respond to calls for change, or to adopt new methods (Sharpe, 2013).

The main changes called for by New Statistics advocates include:

1. Understanding the limitations of significance tests.
2. The need to report effect sizes and CIs.
3. Greater use of meta-analysis to summarize effect size information across studies.

All introductory statistics books I know of cover statistical significance tests and CIs, and most discuss effect size. Adopting the New Statistics perspective does not require you to learn anything new. New Statistics advocates only ask you to think about topics such as statistical significance tests from a more critical perspective. Even though you have probably studied CIs and effect size before, review can be enlightening. This chapter also includes a brief introduction to meta-analysis.

1.3 COMMON MISINTERPRETATIONS OF *p* VALUES

Advocates of the New Statistics have pointed out that misunderstandings about interpretation of p values are widespread. In a survey of researchers that asked which statements about p values they believed to be correct, large numbers of them endorsed incorrect interpretations (Mittag &

Thompson, 2000). Statistics education needs to be improved so that people who use NHST understand its limitations.

There are numerous problems with *p* values that lead to misunderstandings.

1. A *p* value cannot tell us what we want to know. We would like to know, on the basis of our data, something about the likelihood that a research hypothesis (usually an alternative hypothesis) is true. Instead, a *p* value tells us, often very inaccurately, about the probability of obtaining the values of *M* and *t* we found using our sample data, given that the null hypothesis is correct (Cohen, 1994).
2. Common practices, such as running multiple tests and selecting only a few to report on the basis of small *p* values, make *p* values very inaccurate information about risk for Type I decision error.
3. Even if we follow the rules and do everything "right," there will always be risk for decision error. Ideal descriptions of NHST require us to obtain a random sample from the population of interest, satisfy all the assumptions for the test statistic, have no problems with missing values or outliers, do one significance test, and then stop. Even if we could do this (and usually we can't), there would still be nonzero risks for both Type I and Type II decision errors. Because of sampling error, there is an intrinsic uncertainty that we cannot get rid of.
4. There is a fairly common misunderstanding that *p* values tell us something about the size, strength, or importance of an effect. Published papers sometimes include statements like "with $p < .001$, the effect was highly significant." In everyday language, *significant* means important, large, or worthy of notice. However, small *p* values can be obtained even for trivial effects if sample *N* is large enough. We need to distinguish between *p* values and effect size. Chapter 9 in Volume I (Warner, 2020) discusses this further.

From Volume I (Warner, 2020), here are examples of some things you should not say about *p* values. A more complete list of misconceptions to avoid is provided by Greenland et al. (2016).

Never make any of the following statements:

- $p = .000$ (the risk for Type I error can become very small, but in theory, it is never 0).
- *p* was "highly" significant. This leads readers to think that your effect was "significant" in the way we define *significant* in everyday language: large, important, or worthy of notice. Other kinds of effect size information (not *p* values) are required to evaluate the practical or clinical significance of the outcome of a study.
- *p* was "almost" significant (or synonymous terms such as *close to* or *marginally* significant). This language will make people who use NHST in traditional ways, and New Statistics advocates, cringe.
- For "small" *p* values, such as $p = .04$, we cannot say:

  Results were not due to chance or could not be explained by chance. (We cannot know that!)

  Results are likely to replicate in future studies.

  The null hypothesis (H_0) is false.

  We accept (or have proved) the alternative hypothesis.

We also cannot use (1 – *p*), for example (1 – .04) = .96, to make probability statements such as:

> There is a 96% chance that results will replicate.
>
> There is a 96% chance that the null hypothesis is false.

- For *p* values larger than .05, we cannot say, "Accept the null hypothesis."

The language we use to report results should not overstate the strength of the evidence, imply large effect sizes in the absence of careful evaluation of effect size, overgeneralize the findings, or imply causality when rival explanations cannot be ruled out. We should never say, "This study proves that. . . ." Any one study has limitations. As suggested in Volume I (Warner, 2020): It is better to think about research in terms of degrees of belief. As we obtain additional high-quality evidence, we may become more confident of a belief. If high-quality inconsistent evidence arises, that should make us rethink our beliefs.

We can say things such as:

- The evidence in this study is consistent with the hypothesis that . . .
- The evidence in this study was not consistent with the hypothesis that . . .

Hypothesis can be replaced by similar terms, such as *prediction*.

Misunderstandings of *p* values, and what they can and cannot tell us, have been one of several contributing factors in a "replication crisis."

1.4 PROBLEMS WITH NHST LOGIC

The version of NHST presented in statistics textbooks and used by many researchers in social and behavioral science is an amalgamation of ideas developed by Fisher, Neyman, and Pearson (Lenhard, 2006). Neyman and Pearson strongly disagreed with important aspects of Fisher's thinking, and probably none of them would endorse current NHST logic and practices. Here are some commonly identified concerns about NHST logic.

1. **NHST turns an uncertainty continuum into a true/false decision.** Cohen (1994) and Rosnow and Rosenthal (1989) argued that we should think in terms of a continuum of likelihood:

   > A successful piece of research doesn't conclusively settle an issue, it just makes some theoretical proposition to some degree more likely. . . . How much more likely this single research makes the proposition depends on many things, but not on whether *p* is equal to or greater than .05: .05 is not a cliff but a convenient reference point along the possibility-probability continuum. (Cohen, 1994)
   >
   > Surely, God loves the .06 nearly as much as the .05. (Rosnow & Rosenthal, 1989)
   >
   > One way to avoid treating .05 as a cliff is to report "exact" *p* values, as recommended by the American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson & Task Force on Statistical Inference, APA Board of Scientific Affairs, 1999). The APA recommended that authors report "exact" values, such as $p = .032$, instead of a yes/no judgment of whether a result is significant or nonsignificant on the basis of $p < .05$ or $p > .05$. The possibly

annoying quotation marks for "exact" are meant as a reminder that in practice, obtained *p* values often seriously underestimate the true risk for Type I error.

2. **NHST cannot tell us what we want to know.** We would like to know something like the probability that our research or alternative hypothesis is true, or the probability that the finding will replicate in future research, or how strong the effects were. In fact, NHST can tell us only the (theoretical) probability of obtaining the results in our data, given that H_0 is true (Cohen, 1994). NHST does not even do this well, given problems with its use in actual research practice.
3. Some philosophers of science argue that **progress in science requires us to discard faulty or incorrect evidence**. However, when researchers reject H_0, this is not "falsification" in that sense.[1]
4. **NHST is trivial because H_0 is always false.** Any nonzero difference (between μ_1 and μ_2) can be judged statistically significant if the sample size is sufficiently large (Kline, 2013).
5. **NHST requires us to think in terms of double negatives** (and people aren't very good at understanding double negatives). First, we set up a null hypothesis (of no treatment effect) that we almost always do not believe, and then we try to obtain evidence that would lead us to doubt this hypothesis. Double negatives are confusing and inconsistent with every day "psycho-logic" (Abelson & Rosenberg, 1958). In everyday reasoning, people have a strong preference to seek confirmatory evidence. People (including researchers) are confused by double negatives.
6. **NHST is misused in many research situations.** Assumptions and rules for proper use of NHST are stringent and are often violated in practice (as discussed in the next two sections). These violations often invalidate the inferences people want to make from *p* values.

Despite these criticisms, an argument can be made that NHST serves a valuable purpose when it is not misused. It can help assess whether results obtained in a study would be likely or unlikely to occur just because of sampling error when H_0 is true (Abelson, 1997; Garcia-Pérez, 2017). However, information about sampling error is also provided by CIs, in a form that may be less likely to lead to misunderstanding and yes/no thinking (Cumming, 2012).

1.5 COMMON MISUSES OF NHST

In actual practice, applications of NHST often do not conform to the ideal requirements for their use. Three sets of conditions are important for the proper use of NHST. I describe these as assumptions, rules, and handling of specific problems such as outliers. (These are fuzzy distinctions.)

In actual practice, it is difficult to satisfy all the requirements for *p* to be an accurate estimate of risk for Type I error. When these requirements are not met, values of *p* that appear in computer program results are biased; usually they underestimate the true risk for Type I error. When the true risk for Type I error is underestimated, both readers and writers of research reports may be overconfident that studies provide support for claims about findings. This can lead to publication and press-release distribution of false-positive results (Woloshin, Schwartz, Casella, Kennedy, & Larson, 2009). Inconsistent and even contradictory media reports of research findings may erode public trust and respect for science.

1.5.1 Violations of Assumptions

Most statistics textbooks precede the discussion of each new statistic with a list of formal mathematical assumptions about distribution shapes, independence of observations and residuals, and so forth. The list of assumptions for parametric analyses such as the independent-samples *t* test and one-way between-*S* ANOVA include:

- Data on quantitative variables are assumed to be normally distributed in the population from which samples were randomly drawn.
- Variances of scores in populations from which samples for groups were randomly drawn are assumed to be equal across groups (the homogeneity of variance assumption)
- Observations must be independent of one another. (Some textbooks do not explain this very important assumption clearly. See Chapter 2 in Volume I [Warner, 2020].)

For Pearson's *r* and bivariate regression, additional assumptions include:

- The relation between *X* and *Y* is linear.
- The variances of *Y* scores at each level of *X* are equal.
- Residuals from regression are uncorrelated with one another.

Advanced analyses often require additional assumptions.

Textbooks often provide information about evaluation of assumptions. However, most introductory data analysis exercises do not require students to detect or remedy violations of assumptions. The need for preliminary data screening and procedures for screening aren't clear in most introductory books. For NHST results to be valid, we need to evaluate whether assumptions are violated. However, journal articles often do not report whether assumptions were evaluated and whether remedies for violations were applied (Hoekstra, Kiers, & Johnson, 2012).

1.5.2 Violations of Rules for Use of NHST

I use the term *rules* to refer to other important guidelines about proper use of NHST. These are not generally included in lists of formal assumptions about distribution and independence of observations. These rules are often implicit; however, they are very important. These include the following:

- **Select the sample randomly from the (actual) population of interest** (Knapp, 2017). This is important whether you think about NHST in the traditional or classic manner, as a way to answer a yes/no question about the null hypothesis, or in terms of the New Statistics, with greater focus on CIs and less focus on *p* values. Bad practices in sampling limit generalizability of results and also compromise the logic of procedures of NHST.
- In practice, researchers often use convenience samples. When they want to generalize results, they imagine hypothetical populations similar to the sample in the study (invoking the idea of "proximal similarity" [Trochim, 2006] as justification for generalization beyond the sample). The use of convenience samples does not correspond closely to the situations the original developers of inferential statistics had in mind. For example, in industrial quality control, a population could be all the objects made by a factory in a month; the sample could be a random subset of

these objects. The logic of NHST inferential statistics makes more sense for random sampling. Studies based on accidental or convenience samples create much more difficult inference problems.

- **Select the statistical test and criterion for statistical significance (e.g., $\alpha < .05$, two tailed) prior to analysis.** This is important if you want to interpret *p* values as they have often been interpreted in the past, as a basis to make a yes/no decision about a null hypothesis. This rule is often violated in practice. For example, data analysts may use asterisks that appear next to correlations in tables and report that for one asterisk, $p < .05$; for two asterisks, $p < .01$; and for three asterisks, $p < .001$. Using asterisks to report a significance level separately for each correlation could be seen as implicitly setting the α criterion after the fact. On the other hand, many authorities recommend that instead of selecting specific α criteria, you should report an exact *p* value and not use the *p* value to make a yes/no decision about the believability of the null hypothesis. In other words, do not use *p* values as the basis to make statements such as "the result was statistically significant" or "reject H_0." Advocates of the New Statistics recommend that we should not rely on *p* values to make yes/no decisions.
- **Perform only one significance test** (or at most a small number of tests). The opposite of this is: Perform numerous statistical tests, and/or numerous variations of the same basic analysis, and then report only a few "statistically significant" results. This practice is often called *p*-hacking. Other names for *p*-hacking include data fishing, "the garden of forking paths" (Gelman & Loken, 2013), or my personal favorite, torturing the data until they confess (Mills, 1993).

Introductory statistics books usually discuss the problem of inflated risk for Type I error in the context of post hoc tests for ANOVA. They do not always make it clear that this problem is even more serious when people run dozens or hundreds of *t* tests or correlations.

1.5.3 Ignoring Artifacts and Data Problems That Bias *p* Values

Many artifacts that commonly appear in real data influence the magnitude of parameter estimates (such as *M*, *SD*, *r*, and *b*, among others) and *p* values. These include, but are not limited to:

- Univariate, bivariate, and multivariate outliers.
- Missing data that are not missing randomly.
- Measurement problems such as unreliability. For example, the obtained value of r_{xy} is attenuated (reduced) by unreliability of measures for *X* and *Y*.
- Mismatch of distribution shapes (for Pearson's *r* and regression statistics) that constrain the range of possible *r* values.

1.5.4 Summary

Consider an *F* ratio in a one-way between-*S* ANOVA. The logic for NHST goes something like this: If we formulate hypotheses and establish criteria for statistical significance and sample size prior to data collection, and if the null hypothesis is true, and if we take a random sample from the population of interest, and if all assumptions for the statistic are satisfied, and if we have not broken important rules for proper use of NHST, and if there are no artifacts such as outliers and missing values, then *p* should be an unbiased estimate of the likelihood

of obtaining a value of *F* as large as, or larger than, the *F* ratio we obtained from our data. (Additional ifs could be added in many situations.)

This is a long conditional statement. The point is: Values of statistics such as *F* and *p* can provide the information described in ideal or imaginary situations in textbooks only when all of these conditions are satisfied. In actual research, one or many of these assumptions about conditions are violated. Therefore, statistics such as *F* and *p* rarely provide a firm basis for the conclusions described for ideal or imaginary research situations in textbooks. Problems with any of these (assumptions, rules, and artifacts) can result in biased *p* values that in turn may lead to false-positive decisions.

In real-life applications of statistics, it may be impossible to avoid all these problems. For all these reasons, I suggest that most *p* values should be taken with a very large grain of salt. *P* values are least likely to be misleading in simple experiments with a limited number of analyses, such as ANOVA with post hoc tests. They are highly likely to be misleading in studies that include large numbers of variables that are combined in different ways using many different analyses.

It is difficult to prioritize these problems; my guess is that violations of rules (such as running large numbers of significance tests and *p*-hacking) and neglect of sources of artifact (such as outliers) often create greater problems with *p* values in practice than violations of some of the formal assumptions about distributions of scores in populations (such as homogeneity of variance).

It requires some adjustment in thinking to realize that, to a very great extent, the numbers we obtain at the end of an analysis are strongly influenced by decisions made during data collection and analysis (Volume I [Warner, 2020]). Beginning students may think that final numerical results represent some "truth" about the world. We need to understand that with different data analysis decisions, we could have ended up with quite different answers. Greater transparency in reporting (Simmons, Nelson, & Simonsohn, 2011) helps readers understand the degree to which results may have been influenced by a data analyst's decisions.

1.6 THE REPLICATION CRISIS

Misuse and misinterpretation of statistics (particularly *p* values) is one of many factors that has contributed to rising concerns about the reproducibility of high-profile research findings in psychology. To evaluate reproducibility of research results, Brian Nosek and Jeff Spies founded the Center for Open Science in 2013 (Open Science Collaboration, 2015). Their aim was to increase openness, integrity, and reproducibility of scientific research. Participating scientists come from many fields, including astronomy, biology, chemistry, computer science, education, engineering, neuroscience, and psychology. Results reported for the first group of studies evaluated were disturbing. They conducted replications of 100 studies (both correlational and experimental) published in three psychology journals, using large samples (to provide adequate statistical power) and original materials if available. The average effect sizes were about half as large as the original results. Only 39 of the 100 replications yielded statistically significant outcomes (all original studies were "statistically significant"). This was not quite as bad as it sounds, because many original effect sizes associated with nonsignificant outcomes were within 95% CIs on the basis of replication effect sizes (Baker, 2015; Open Science Collaboration, 2015). These results attracted substantial attention and concern.

Failures to replicate have also been noted in biomedical research. Ioannidis (2005) examined 49 highly regarded medical studies from 13 prior years. He compared initial claims for intervention effectiveness with results in later studies with larger samples; 7 (16%) of the original studies were contradicted, and another 7 (16%) had smaller effects than the original study. Later studies have yielded even less favorable results. Begley and Ellis (2012)

reported that biotechnology firm Amgen tried to confirm results from 53 landmark studies about issues such as new approaches to targeting cancers and alternative clinical uses for existing therapeutics. Findings were confirmed for only 6 (11%) studies. Baker and Dolgin (2017) noted that early results from the Cancer Reproducibility Project's examination of 6 cancer biology studies were mixed.

Do these replication failures indicate a "crisis"? That is debatable. Only a small subset of published studies were tested. Some of the original studies were chosen for replication because they reported surprising or counterintuitive results. Examination of *p* values is not the best way to assess whether results have been reasonably well replicated; *p* values are "fickle" and difficult to reproduce (Halsey, Curran-Everett, Vowler, & Drummond, 2015). It may be better to evaluate reproducibility using effect sizes or CIs instead of *p* values. Critics of the reproducibility projects argue that the replication methods and analyses were flawed (Gilbert, King, Pettigrew, & Wilson, 2016). It would be premature to conclude that large proportions of all past published research results would not replicate; however, concerns raised by failures to replicate should be taken seriously.

A failure to reproduce results does not necessarily mean that the original or past study was wrong. The replication study may be flawed, or the results may be context dependent (and might appear only in the specific circumstances in an earlier study, and not under the conditions in the replication study).

Concerns about reproducibility have led to a call for new approaches to reporting results, often called the New Statistics, along with a movement toward preregistration of study plans and Open Science, in which researchers more fully share information about study design and statistical analyses.

Many changes in research practice will be needed to improve reproducibility of research results (Wicherts et al., 2016). Misuse and misinterpretation of statistical significance tests (and *p* values) to make yes/no decisions about whether studies are "successful" have contributed to problems in replication. Some have even argued that NHST and *p* values are an inherently flawed approach to evaluation of research results (Krueger, 2001; Rozeboom, 1960). Cumming (2014) and others argue that a shift in emphasis (away from statistical significance tests and toward reports of effect size, CIs, and meta-analysis) is needed. However, many published papers still do not include effect size and CIs for important results (Watson, Lenz, Schmit, & Schmit, 2016).

1.7 SOME PROPOSED REMEDIES FOR PROBLEMS WITH NHST

1.7.1 Bayesian Statistics

Some authorities argue that we got off on the wrong foot (so to speak) when we adopted NHST in the early 20th century. Probability is a basic concept in statistical significance testing. The examples used to explain probability suggest that it is a simple concept. For example, if you draw 1 card at random from a deck of 52 cards with equal numbers of diamonds, hearts, spades, and clubs, what is the probability that the card will be a diamond? This example does not even begin to convey how complicated the notion of probability becomes in more complex situations (such as inference from sample to population).

NHST is based on a "frequentist" understanding of probability; this is not the only possible way to think about probability, and other approaches (such as Bayesian) may work better for some research problems. A full discussion of this problem is beyond the scope of this chapter; see Kruschke and Liddell (2018), Little (2006), Malakoff (1999), or Williamson (2013).

Researchers in a few areas of psychology use Bayesian methods. However, students typically receive little training in these methods. Whatever benefits this might have, a major shift toward the use of Bayesian methods in behavioral or social sciences seems unlikely to happen any time soon.

1.7.2 Replace α = .05 with α = .005

It has recently been suggested that problems with NHST could be reduced by setting the conventional α criterion to .005 instead of the current .05 (Benjamin et al., 2017). This would establish a more stringent standard for announcement of "new" findings. However, given the small effect sizes in many research areas, enormous sample sizes would be needed to have reasonable statistical power with α = .005. This would be prohibitively costly. Bates (2017) and Schimmack (2017) argued that this approach is neither necessary nor sufficient and that it would make replication efforts even more unlikely. A change to this smaller α level is unlikely to be widely adopted.

1.7.3 Less Emphasis on NHST

The "new" statistics advocated by Cumming (2012, 2014) calls for a shift of focus. He recommended that research reports should focus more on

- confidence intervals,
- effect size information, and
- meta-analysis to combine effect size information across studies.

How "new" is the New Statistics? As noted by Cumming (2012) and others, experts have been calling for these changes for more than 40 years (e.g., Morrison & Henkel, 1970; Cohen, 1990, 1994; Wilkinson & Task Force on Statistical Inference, APA Board of Scientific Affairs, 1999). Cumming (2012, 2014) bolstered these arguments with further discussion of the ways that CIs (vs. *p* values) may lead data analysts to think about their data. Some argue that the New Statistics is not really "new" (Palij, 2012; Savalei & Dunn, 2011); CIs and significance tests are based upon the same information about sampling error. In practice, many readers may choose to convert CIs into *p* values so that they can think about them in more familiar terms. However, effect size reporting is critical; it provides information that is not obvious from examination of *p* values.

Unlike a shift to Bayesian approaches, or the use of α = .005, including CIs and effect sizes in research reports would not be difficult or costly. In general, researchers have been slow to adopt these recommendations (Sharpe, 2013). The *Journal of Basic and Applied Social Psychology* (Trafimow & Marks, 2015) now prohibits publication of *p* values and related NHST results.

The following sections review the major elements of the New Statistics: CIs and effect size. CIs and effect size are both discussed in Volume I (Warner, 2020) for each bivariate statistic. A brief introduction to meta-analysis is also provided.

1.8 REVIEW OF CONFIDENCE INTERVALS

A confidence interval is an interval estimate for some unknown population characteristic or parameter (such as μ, the population mean) based on information from a sample (such as *M*, *SD*, and *N*). CIs can be set up for basic bivariate statistics using simple formulas. Unfortunately SPSS does not provide CIs for some statistics, such as Pearson's *r*. For more advanced statistics, CIs can be set up using methods such as bootstrapping, which is discussed in Chapter 15, on structural equation modeling, later in this book.

1.8.1 Review: Setting Up CIs

Consider an example of the CI for one sample mean, *M*. Suppose a data analyst has IQ scores for a sample of $N = 100$ cases, with these sample estimates: $M = 105$, $SD = 15$. In addition

to reporting that mean IQ in the sample was $M = 105$, an interval estimate (a 95% CI) can be constructed, with lower and upper boundaries. The procedure used in this example can be used only when the sample statistic is known to have a normally shaped sampling distribution and when N is large enough that the standard normal or z distribution can be used to figure out what range of values lies within the center 95% of the distribution. (With smaller samples, t distributions are usually used.)

These are the steps to set up a CI:

- Decide on ***C* (level of confidence)** (usually this is 95%).
- Assuming that your sample statistic has a normally shaped sampling distribution, use the "critical values" from a z or standard normal distribution that correspond to the middle 95% of values. For a standard normal distribution, the middle 95% corresponds to the interval between $z_{lower} = -1.96$ and $z_{upper} = +1.96$. (Rounding these z values to –2 and +2 is reasonable when thinking about estimates.)
- Find the standard error (*SE*) for the sample statistic. The *SE* depends on sample size and standard deviation. For a sample mean, $SE_M = SD/\sqrt{N}$. Other sample statistics (such as r, b, and so forth) also have *SE*s that can be estimated.
- On the basis of $SD = 15$, and $N = 100$, we can compute the standard error of the sampling distribution for M: $SE_M = 15/\sqrt{100} = 15/10 = 1.5$.
- Now we combine SE_M with M and the z critical values that correspond to the middle 95% of the standard normal distribution to compute the CI limits:

$$\text{Lower limit} = M + z_{lower} \times SE_M = 105 - 1.96 * 1.5 = 105 - 2.94 = 102.06.$$

$$\text{Upper limit} = M + z_{upper} \times SE_M = 105 + 1.96 * 1.5 = 105 + 2.94 = 107.94.$$

This would be reported as "95% CI [102.06, 107.94]."

This procedure can be generalized and used with many other (but not all) sample statistics. To use this procedure, an estimate of the value of $SE_{statistic}$ is needed, and the sampling distribution for the statistic must be normal:

$$\text{Lower limit} = \text{Statistic} + z_{lower} \times SE_{statistic.} \quad (1.1)$$

$$\text{Upper limit} = \text{Statistic} + z_{upper} \times SE_{statistic.} \quad (1.2)$$

The statistic can be $(M_1 - M_2)$, r, or a raw-score regression slope b, for example. In more advanced analyses such as structural equation modeling, it is sometimes not possible to calculate the *SE* values for path coefficients directly, and it may be unrealistic to expect sampling distributions to be normal in shape. In these situations, Equations 1.1 and 1.2 cannot be used to set up CIs. The chapters that introduce structural equation modeling and logistic regression discuss different procedures to set up CIs for these situations.

1.8.2 Interpretation of CIs

It is incorrect to say that there is a 95% probability that the true population mean μ lies within a 95% CI. (It either does, or it doesn't, and we cannot know which.) We can make a long-range prediction that, if we have a population with known mean and standard deviation,

and set a fixed sample size, and draw thousands of random samples from that population, that 95% of the CIs set up using this information will contain μ and the other 5% will not contain μ. Cumming and Finch (2005) provided other correct interpretations for CIs.

1.8.3 Graphing CIs

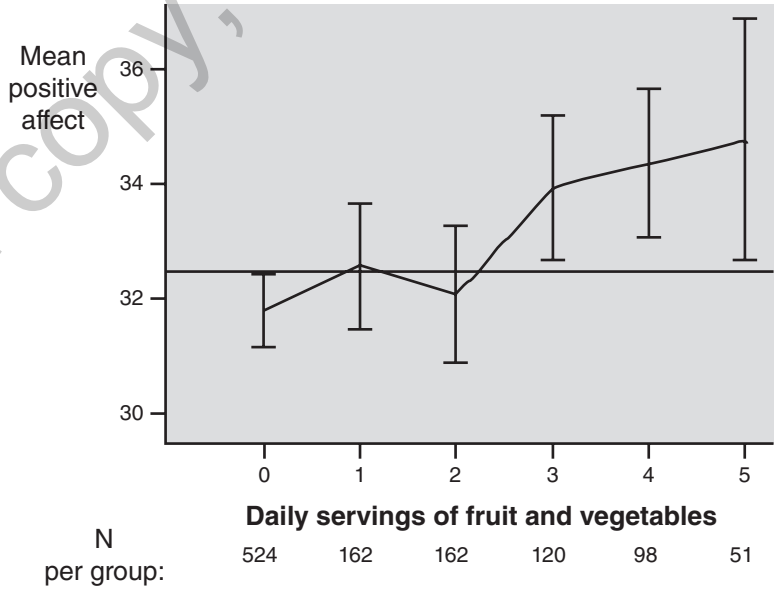
Upper and lower limits of CIs may be reported in text, tables, or graphs. One common type of graph is an error bar chart, as shown in Figure 1.1. (Bar charts can also be set up with error bars.) For either error bar or bar chart graphs, the graph may be rotated, such that error bars run from left to right instead of from bottom to top.

The data in Figure 1.1 are excerpted from an actual study. Undergraduates reported positive affect and the number of servings of fruit and vegetables they consumed in a typical day. Earlier research suggested that higher fruit and vegetable intake was associated with higher positive affect. Given the large sample size, number of servings could be treated as a group variable (i.e., the first group ate no servings of fruits and vegetables per day, the second group ate one serving per day, etc.) This was useful because past research suggested that the increase in positive affect might not be linear.

The vertical "whiskers" in Figure 1.1 show the 95% CI limits for each group mean. The horizontal line that crosses the *Y* axis at about 32.4 helps clarify that the CI for the zero servings of fruits and vegetables group did not overlap with the CIs for the groups of persons who ate three, four, or five servings per day.

In graphs of this type, the author must indicate whether the error bars correspond to a CI (and what level of confidence). Some graphs use similar-looking error bar markers to indicate the interval between –1 SE_M and +1 SE_M or the interval between –1 *SD* and +1 *SD*.

Figure 1.1 Mean Positive Affect for Groups With Different Fruit and Vegetable Intake (With 95% CI Error Bars)



Source: Adapted from Warner, Frye, Morrell, and Carey (2017).

1.8.4 Understanding Error Bar Graphs

A reader can make two kinds of inferences from error bars in this type of graph (Figure 1.1). First, error bars can be used to guess which group means differed significantly. Cumming (2012, 2014) cautioned that analysts should not automatically convert CI information into *p* values for significance tests when they think about their results. However, if readers choose to do that, it is important to understand the way CIs and two-tailed *p* values are related. In general, if the CIs for two group means do not overlap in graphs such as Figure 1.1, the difference between means is statistically significant (assuming that the level of confidence corresponds to the α level, i.e., 95% confidence and $\alpha = .05$, two tailed). On the other hand, the difference between a pair of group means can be statistically significant even if the CIs for the means overlap slightly. Whether the difference is statistically significant depends on the amount of overlap between CIs (Cumming & Finch, 2005; Knezevic, 2008).

The nonoverlapping CIs for the zero-servings group and five-servings group indicates that if a *t* test were done to compare these two group means, using $\alpha = .05$, two tailed, this difference would be statistically significant. There is some overlap in the CIs for the two-servings and three-servings groups. This difference might or might not be statistically significant using $\alpha = .05$, two tailed.

The second kind of information a reader should look for is practical or clinical significance. Mean positive affect was about 34 for the five-servings group and 32 for the zero-servings group. Is that difference large enough to value or care about? Would a typical person be motivated to raise fruit and vegetable consumption from zero to five servings if that meant a chance to increase positive affect by two points? (Maybe there are easier ways to "get happy.")

Numbers on the scale for positive affect scores are meaningless unless some context is provided. In this example, the minimum possible score for positive affect was 10 points, and the maximum was 50 points. A 2-point difference on a 50-point rating scale does not seem like very much. Also note that this graph "lies with statistics" in a way that is very common in both research reports and the mass media. The *Y* axis begins at about 30 points rather than the actual minimum value of 10 points. How different would this graph look if the *Y* axis included the entire possible range of values from 10 to 50?

In the final analyses in our paper (Warner et al., 2017), fruit and vegetable intake uniquely predicted about 2% of the variance in positive affect after controlling for numerous other variables that included exercise and sleep quality. That 2% was statistically significant. However, on the basis of 2% of the variance and a two-point difference in positive affect ratings for the low versus high fruit and vegetable consumption groups, I would not issue a press release urging people to eat fruit and get happy. Other variables (such as gratitude) have much stronger associations with positive affect. (It may be of theoretical interest that consumption of fruits and vegetables, but not sugar or fat consumption, was related to positive affect. Fruit and vegetable consumption is related to other important outcomes such as physical health.)

The point is: Information about actual and potential range of scores for the outcome variable can provide context for interpretation of scores (even when they are in essentially meaningless units). Readers also need to remember that the selection of a limited range of values to include on the *Y* axis creates an exaggerated perception of group differences.

1.8.5 Why Report CIs Instead of, or in Addition to, Significance Tests?

Cumming (2012) and others suggest these possible advantages of focusing on CIs rather than *p* values:

1. **Reporting the CI can move us away from the yes/no thinking** involved in statistical significance tests (unless we use the CI only to reconstruct the statistical significance test).

2. **CIs make us aware of the lack of precision of our estimates** (of values such as means). Information about lack of precision is more compelling when scores on a predicted variable are in meaningful units. Consider systolic blood pressure, given in millimeters of mercury (mm Hg). If the 95% CI for systolic blood pressure in a group of drug-treated patients ranges from 115 mm Hg (not considered hypertensive) to 150 mm Hg (hypertensive), potential users of the drug will be able to see that mean outcomes are not very predictable. (On the other hand, if the CI ranges from 115 to 120 mm Hg, mean outcomes can be predicted more accurately.)
3. **CIs may be more stable across studies than *p* values.** In studies of replication and reproducibility, overlap of CIs across studies may be a better way to assess consistency than asking if studies yield the same result on the binary outcome judgment: significant or not significant. *P* values are "fickle"; they tend to vary across samples (Halsey et al., 2015). Asendorpf et al. (2013) recommended that evaluation of whether two studies produce consistent results should focus on CI overlap rather than on "vote counting" (i.e., noticing whether both studies had $p < .05$).

Data analysts hope that CIs will be relatively narrow, because if they are not, it indicates that estimates of mean have considerable sampling error. Other factors being equal, the width of a CI depends on these factors:

- As *SD* increases (other factors being equal), the width of the *CI* increases.
- As level of confidence increases (other factors being equal), the width of the *CI* increases.
- As *N* increases (other factors being equal), the width of the *CI* decreases.

Despite calls to include CIs in research reports, many authors still do not do so (Sharpe, 2013). This might be partly because, as Cohen (1994) noted, they are often "so embarrassingly large!"

1.9 EFFECT SIZE

Bivariate statistics introduced in Volume I (Warner, 2020) were accompanied by a discussion of one (or sometimes more than one) effect size indexes. For χ^2, effect sizes include Cramer's *V* and ϕ**.** Pearson's *r* and r^2 directly provide effect size information. For statistics such as the independent-samples *t* test, several effect sizes can be used; these include point biserial *r* (r_{pb}), **Cohen's *d***, η, and η^2. It is also possible to think about the ($M_1 - M_2$) difference as information about practical or clinical effect size terms if the dependent variable is measured in meaningful units such as dollars, kilograms, or inches. For ANOVA, η and η^2 are commonly used. Rosnow and Rosenthal (2003) discussed additional, less widely used effect size indexes.

1.9.1 Generalizations About Effect Sizes

1. Effect size is independent of sample size. For example, the magnitude of Pearson's *r* does not systematically increase as *N* increases.[2]
2. Some effect sizes have a fixed range of possible values (*r* ranges from –1 to +1), but other effect sizes do not (Cohen's *d* is rarely higher than 3 in absolute value, but it does not have a fixed limit).

3. Many effect sizes are in unit-free (or standardized) terms. For example, the magnitude of Pearson's r is not related to the units in which X and Y are measured.
4. On the other hand, effect size information can be presented in terms of the original units of measurement (e.g., $M_1 - M_2$). This is useful when original units of measurement were meaningful (Pek & Flora, 2018).
5. Some effect sizes can be directly converted (at least approximately) into other effect sizes (Rosnow & Rosenthal, 2003).
6. Cohen's (1988) guidelines for verbal labeling of effect sizes are widely used; these appear in Table 1.1. Alternative guidelines based on Fritz, Morris, and Richler (2012) appear in Table 1.2.
7. The value of a test statistic (such as the independent-samples t test) depends on both effect size and sample size or df. This is explained further in the next section.
8. Many journals now call for reporting of effect size information. However, many published research reports still do not include this information.
9. Judgments about the clinical or practical importance of research results should be based on effect size information, not based on p values (Sullivan & Feinn, 2012).
10. If you read a journal article that does not include effect size information, there is usually enough information for you to compute an effect size yourself. (There should be!)
11. Computer programs such as SPSS often do not provide effect sizes; however, effect sizes can be computed from the information provided.

Table 1.1 Suggested Verbal Labels for Cohen's *d* and Other Common Effect Sizes

Verbal Label Suggested by Cohen (1988)	Cohen's *d*	*r*, r_{pb},[a] *b*, Partial *r*, *R*, or β	r^2, R^2, or η^2
Large effect	0.8	.371	.138
(In-between area)	0.7	.330	.109
	0.6	.287	.083
Medium effect	0.5	.243	.059
(In-between area)	0.4	.196	.038
	0.3	.148	.022
Small effect	0.2	.100	.010
(In-between area)	0.1	.050	.002
No effect	0.0	.000	.000

Source: Adapted from Cohen (1988).

a. Point biserial r is denoted r_{pb}. For an independent-samples t test, r_{pb} is the Pearson's r between the dichotomous variable that represents group membership and the Y quantitative dependent variable.

Table 1.2 Effect Size Interpretations

Research Question	Effect Sizes	Minimum Reportable Effect[a]	Moderate Effect	Large Effect
Difference between two group means	Cohen's *d*	.41	1.15	2.70
Strength of association: linear	r, r_{pb}, R, partial r, β, tau	.2	.5	.8
Squared linear association estimates	r^2, partial r^2, R^2, adjusted R^2, sr^2	.04	.25	.64
Squared association (not necessarily linear)	η^2 and partial η^2	.04	.25	.64
Risk estimates[b]	RR, OR	2.0	3.0	4.0

Source: Adapted from Fritz et al. (2012).

a. The minimum values suggested by Fritz et al. are much higher than the ones proposed by Cohen (1988).

b. Analyses such as logistic regression (in which the dependent variable is a group membership, such as alive vs. dead) provide information about relative or comparative risk, for example, how much more likely is a smoker to die than a nonsmoker? This may be in the form of relative risk (RR) and an odds ratio (OR). See Chapter 16.

12. In the upcoming discussion of meta-analysis, examples often focus on effect sizes such as Cohen's *d* that describe the difference between group means for treatment and control groups. However, raw or standardized regression slope coefficients can also be treated as effect sizes in meta-analysis (Nieminen, Lehtiniemi, Vähäkangas, Huusko, & Rautio, 2013; Peterson & Brown, 2005).
13. CIs can be set up for many effect size estimates (Kline, 2013; Thompson, 2002b). Ultimately, it would be desirable to report these along with effect size. In the short term, just getting everyone to report effect size for primary results is probably a more reasonable goal.

1.9.2 Test Statistics Depend on Effect Size Combined With Sample Size

Consider the independent-samples *t* test. M_1 and M_2 denote the group means, SD_1 and SD_2 are the group standard deviations, and n_1 and n_2 denote the number of cases in each group. One of the effect sizes used with the independent-samples *t* is Cohen's *d* (the standardized distance or difference between the sample means M_1 and M_2). The difference between the sample means is standardized (converted to a unit-free distance) by dividing $(M_1 - M_2)$ by the pooled standard deviation s_p:

$$Cohen's\ d = \frac{M_1 - M_2}{s_p}. \tag{1.3}$$

Formulas for s_p sometimes appear complicated; however, s_p is just the weighted average of SD_1 and SD_2, weighted by sample sizes n_1 and n_2.

Sample size information for the independent-samples *t* test can be given as $(\sqrt{df}/2)$, where $df = (n_1 + n_2) - 2$. The formula for the independent-samples *t* test can be given as a function of effect size *d* and sample size, as shown by Rosenthal and Rosnow (1991):

$$t = d\frac{\sqrt{df}}{2}. \tag{1.4}$$

Examining Equation 1.4 makes it clear that if effect size *d* is held constant, the absolute value of *t* increases as the *df* (sample size) increases. Thus, even when an effect size such as *d* is extremely small, as long as it is not zero, we can obtain a value of *t* large enough to be judged statistically significant if sample size is made sufficiently large. Conversely, if the sample size given by *df* is held constant, the absolute value of *t* increases as *d* increases. This dependence of magnitude of the test statistic on both effect size and sample size holds for other statistical tests (I have provided only a demonstration for one statistic, not a proof).

This is the important point: A very large value of *t*, and a correspondingly very small value of *p*, can be obtained even when the effect size *d* is extremely small. A small *p* value does not necessarily tell us that the results indicate a large or strong effect (particularly in studies with very large *N*'s).

Furthermore, both the value of *N* and the value of *d* depend on researcher decisions. For an independent-samples *t* test, other factors being equal, *d* often increases when the researcher chooses types of treatments and/or dosages of treatments that cause large differences in the response variable and when the researcher controls within-group error variance through standardization of procedures and recruitment of homogeneous samples. Some undergraduate students became upset when I explained this: "You mean you can make the results turn out any way you want?" Yes, within some limits. When we obtain statistics in samples, such as values of *M* or Cohen's *d* or *p*, these values depend on our design decisions. They are not facts of nature. See Volume I (Warner, 2020), Chapter 12, for further discussion.

1.9.3 Using Effect Size to Evaluate Theoretical Significance

Judgments about theoretical significance are sometimes made on the basis of the magnitude of standardized effect size indexes such as *d* or *r*. One way to think about the importance of research results is to ask, Given the effect size, how much does this variable add to our ability to predict some outcome of interest, or to "explain variance"? Is the added predictive information sufficient to be "worthwhile" from a theoretical perspective? Is it useful to continue to include this variable in future theories, or are its effects so trivial as to be negligible?

For example, if *X* and *Y* have $r_{xy} = .10$ and therefore, $r^2 = .01$, then only 1% of the variance in *Y* is linearly predictable from *X*. By implication, the other 99% of the variance is related to other variables (or is due to nonlinear associations or is inherently unpredictable). Is it worth expending a lot of energy on further study of a variable that predicts only 1% of the variance? When an effect size is this small, very large *N*'s are needed in future studies in order to have sufficient statistical power (i.e., a reasonably high probability of obtaining a statistically significant outcome). Researchers need to make their own judgments as to whether it is worth pursuing a variable that predicts such a small proportion of variance.

There are two reasons why authors may not report effect sizes. One is that SPSS does not provide effect size information for some common statistics, such as ANOVA. This lack is easy to deal with, because SPSS does provide the information needed to calculate effect size information by hand, and the computations are simple. This information is provided for each statistic in Volume I (Warner, 2020). For example, an η^2 effect size for ANOVA can be obtained by dividing SS_{effect} by SS_{total}. There may be another reason. Cohen (1994) noted that CIs are often embarrassingly large; effect sizes may often be embarrassingly small. It just does not sound very impressive to say, "I have accounted for 1% of the variance."

A long time ago, Mischel (1968) pointed out that correlations between personality measures and behaviors tended to be no larger than $r = .30$. This triggered a crisis and disputes in personality research. Social psychologists argued that the power of situations was much greater than personality. Epstein and O'Brien (1985) argued that it is possible to obtain

higher correlations in personality with broader assessments and that typical effect sizes in social psychology were not much higher. However, at the time, $r = .30$ seemed quite low. This may have been because earlier psychological research in areas such as behavior analysis and psychophysics tended to yield much larger effects (stronger correlations). I wonder whether Cohen's labeling of $r = .3$ as a medium to large effect was based on the observation that in many areas of psychology, effects much larger than this are not common. Nevertheless, accounting for 9% of the variance does not sound impressive.

Prentice and Miller (1992) pointed out that in some situations, even small effects may be impressive. Some behaviors are probably not easy to change, and a study that finds some change in this behavior can be impressive even if the amount of change is small. They cited this example: Physical attractiveness shows strong relationships with some responses (such as interpersonal attraction). It is impressive to note that even in the courtroom, attractiveness has an impact on behavior; unattractive defendants were more likely to be judged guilty and to receive more punishment. If physical attractiveness has effects in even this context, its effects may apply to a very wide range of situations.

Sometimes social and behavioral scientists have effect size envy, imagining that effect sizes in other research domains are probably much larger. In fact, effect sizes in much biomedical research are similar to those in psychology (Ferguson, 2009). Rosnow and Rosenthal (2003) cited an early study that examined whether taking low-dose aspirin could reduce the risk for having a heart attack. Pearson's r (or ϕ) between these two dichotomous variables was $r = .034$. The percentage of men who did not have heart attacks in the aspirin group (51.7%) was significantly higher than the percentage of men who did not have heart attacks in the placebo group (48.3%). Assuming that these results are generalizable to a larger population (and that is always a question), a 3.4% improvement in health outcome applied to 1 million men could translate into prevention of about 34,000 heart attacks. From a public health perspective, $r = .034$ can be seen as a large effect. From the perspective of an individual, the evaluation could be different. An individual might reason, I might change my risk for heart attack from 51.7% (if I do not take aspirin) to 48.3% (if I do take aspirin). From that perspective, the effect of aspirin might appear to be less substantial.

1.9.4 Use of Effect Size to Evaluate Practical or Clinical Importance (or Significance)

It is important to distinguish between statistical significance and practical or clinical significance (Kirk, 1996; Thompson, 2002a). We have clear guidelines how to judge statistical significance (on the basis of p values). What do we mean by clinical or practical significance, and how can we make judgments about this? In everyday use, the word *significant* often means "sufficiently important to be worthy of attention." When research results are reported as statistically significant, readers tend to think that the treatment caused effects large enough to be noticed and valued in everyday life. However, the term *statistically significant* has a specific technical meaning, and as noted in the previous section, a result that is statistically significant at $p < .001$ may not correspond to a large effect size.

For a study comparing group means, practical significance corresponds to differences between group means that are large enough to be valued (a large $M_1 - M_2$ difference). In a regression study, practical significance corresponds to large and "valuable" increases in an outcome variable as scores on the independent variable increase (e.g., a large raw-score regression slope b).

Standardized effect sizes such as Cohen's d are sometimes interpreted in terms of clinical significance. However, examining the difference between group means ($M_1 - M_2$) in their original units of measurement can be a more useful way to evaluate the clinical or practical importance of results (Pek & Flora, 2018). $M_1 - M_2$ provides understandable information

when variables are measured in meaningful and familiar units. Age in years, salary in dollars or euros or other currency units, and body weight in kilograms or pounds are examples of variables in meaningful units. Everyday people can understand results reported in these terms.

For example, if a study that compared final body weight between treatment (1) and control (2) groups, with mean weights M_1 = 153 lb in the treatment group and M_2 = 155 lb in the control group, everyday folks (as well as clinicians) probably would not think that a 2-lb difference is large enough to be noticeable or valuable. Most people would not be very interested in this new treatment, particularly if it is expensive or difficult. On the other hand, if the two group means differed by 20 or 30 lb, probably most people would view that as a substantial difference. Similar comparisons can be made for other different treatment outcomes (such as blood pressure with vs. without drug treatment).

Unfortunately, when people read about new treatments in the media, reports often say that a treatment effect was "statistically significant" or even "highly statistically significant." Those phrases can mislead people to think that the difference between group means (for weight, blood pressure, or other outcomes) in the study was extremely large.

Here are examples of criteria that could be used to judge whether results of studies are clinically or practically significant, that is, whether outcomes are different enough to matter:

- **Are group means so far apart that one mean is above, and the other mean is below, some diagnostic cutoff value?** For example, is systolic blood pressure in a nonhypertensive range for the treatment group and a hypertensive range for a control group?
- **Would people care about an effect this size?** This is relatively easy to judge when the variable is money. Judge and Cable (2004) examined annual salaries for tall versus short persons. They reported these mean annual salaries (in U.S. dollars): tall men, $79,835; short men, $52,704; tall women, $42,425, short women, $32,613. As always in research, there are many reasons we should hesitate to generalize their results to other situations or apply them to ourselves individually. However, tall men earned mean salaries more than $47,000 higher than short women. I am a short woman, and this result certainly got my attention.

In economics, value or "mattering" is called utility. Systematic studies could be done to see what values people (clients, clinicians, and others) attach to specific outcomes. For a person who earns very little money, a $1,000 salary increase may have a lot of value. For a person who earns a lot of money, the same $1,000 increase might be trivial. Utility of specific outcomes might well differ across persons according to characteristics such as age and sex.

- **How large does a difference have to be for most people to even notice or detect it?** At a bare minimum, before we speak of an effect detected in a study as an important finding, it should be noticeable in everyday life (cf. Donlon, 1984; Stricker, 1997).

1.9.5 Uses for Effect Sizes

- Effect sizes should be included in research reports. Standardized effect sizes (such as Cohen's *d* or *r*) provide a basis for labeling strength of relationships between variables as weak, moderate, or strong. Standardized effect sizes can be compared with those found in other studies and in past research. Additional information, such as raw-score regression slopes and group means in original units of measurement, can help readers understand the real-world or clinical implications of findings (at least if the original units of measurement were meaningful).

- Effect size estimates from past research can be used to do statistical power analysis to make sample-size decisions for future research.
- Finally, effect size information can be combined and evaluated across studies using meta-analysis to summarize existing information.

1.10 BRIEF INTRODUCTION TO META-ANALYSIS

A meta-analysis is a summary of effect size information from past research. It involves evaluating the mean and variance of effect sizes combined across past studies. This section provides only a brief overview. For details about meta-analysis, see Borenstein, Hedges, Higgins, and Rothstein (2009) or Field and Gillett (2010).

1.10.1 Information Needed for Meta-Analysis

The following steps are involved in information collection:

1. **Clearly identify the question of interest.** For example, how does number of bystanders (*X*) predict whether a person offers help (*Y*)? What is the difference in mean depression scores (*Y*) between persons who do and do not receive cognitive behavioral therapy (CBT) (*X*)?
2. **Establish criteria for inclusion (vs. exclusion) of studies ahead of time.** Decide which studies to include and exclude. This involves many judgments. Poor-quality studies may be discarded. Studies that are retained must be similar enough in conception and design that comparisons make sense (you can't compare apples and oranges). Reading meta-analyses in your own area of interest can be helpful.
3. **Do a thorough search for past research about this question.** This should include published studies, located using library databases, and unpublished data, obtained through personal contacts.
4. **Create a data file that has at least the following information for each study:**
   a. Author names and year of publication for each study.
   b. Number in sample (and within groups).
   c. Effect size information (you may have to calculate this if it is not provided). The most common effect sizes are Cohen's *d* and *r*. However, other types of effect size may be used.[3]
   d. If applicable, group sizes, means, and standard deviations.
   e. Additional information to characterize studies. If the number of studies included in the meta-analysis is large, it may be possible to analyze these variables as possible "moderators," that is, variables that are related to different effect sizes. In studies of CBT, the magnitude of treatment effect might depend on number of treatment sessions, type of depression, client sex, or even the year when the study was done. There are also "study quality" and study type variables, for example, Was the study double blind or not? Was there a nontreatment control group? Was it a within-*S* or between-*S* design? It is a good idea to have more than one reader code this information and to check for interobserver reliability.

1.10.2 Goals of Meta-Analysis

- **Estimate mean effect size.** When effect sizes are averaged across studies they are usually weighted by sample size (or sometimes by other characteristics of studies).

- **Evaluate the variance of effect sizes across studies.** The variation among effect sizes indicates whether results of studies seem to be homogeneous (that is, they all tended to yield similar effect sizes) or heterogeneous (they yielded different effect sizes). If effect sizes are heterogeneous and the number of studies is reasonably large, a moderator analysis is possible.
- **Evaluate whether certain moderator variables are related to difference in effect sizes.** For example, are smaller effect sizes obtained in recent CBT studies than in those done many years ago?

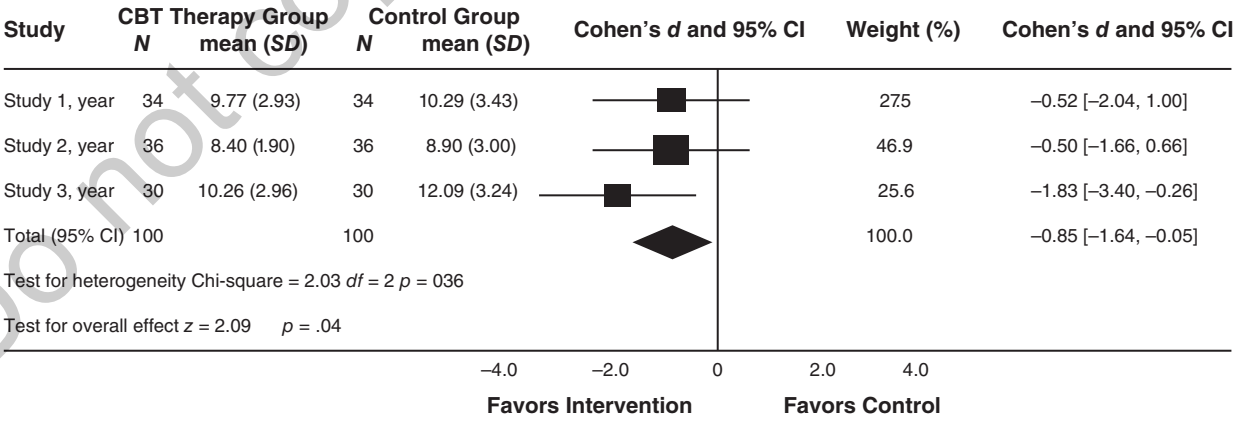
The mechanics of doing a meta-analysis can be complex. For example, the analyst must choose between a fixed- and a random-effects model (for discussion, see Field & Gillett, 2010); a random-effects model is probably more appropriate in many situations. SPSS does not have a built-in meta-analysis procedure; Field and Gillet (2010) provide free downloadable SPSS syntax files on their website, and references to software created by others, including routines in R. See the following sources for guidelines about reporting meta-analysis: Liberati et al. (2009) and Rosenthal (1995).

1.10.3 Graphic Summaries of Meta-Analysis

Forest plots are commonly used to describe results from meta-analysis. Figure 1.2 shows a hypothetical forest plot. Suppose that three studies were done to compare depression scores between a group that has had CBT and a control group that has not had therapy. For each study, the effect size, Cohen's *d*, is the difference between posttest depression scores for the CBT and control groups (divided by the pooled within-group standard deviation). A 95% CI is obtained for Cohen's *d* for each study.

The vertical line down the center of the table is the "line of no effect" that corresponds to $d = 0$. This would be the expected result if population means did not differ between CBT and control conditions. In this example, a negative value of *d* means that the treatment group had a better outcome (i.e., lower depression after treatment) than the control group.

Figure 1.2 Hypothetical Forest Plot for Studies That Assess Posttreatment Depression in Therapy and Control Groups



Source: Adapted with permission from the Royal Australian College of General Practitioners from: Ried K. "Interpreting and understanding meta-analysis graphs: A practical guide." *Australian Family Physician*, 2006; 35(8):635–38. Available at www.racgp.org.au/afp/200608/10624.

Reading across the line for Study 1: Author names and year are provided, then *N*, mean, and *SD* for the CBT and control groups. The horizontal line to the right, with a square in the middle, corresponds to the 95% CI for Cohen's *d* for Study 1. The size of the square is proportional to total *N* for that study. The weight given to information from each study in a meta-analysis can be based on one or more characteristics of studies, such as sample size. The final column provides the exact numerical results that correspond to the graphic version of the 95% CI for Cohen's *d* for each study.

The row denoted "Total" shows the 95% CI for the weighted mean of Cohen's *d* across all three studies, first in graphic and then in numerical form. The "Total" row has a diamond-shaped symbol; the end points of the diamond indicate the 95% CI for the average effect size across studies. This CI did not include 0.

The values in the lower left of the figure answer two questions about the set of effect sizes across all studies. First, does the weighted mean of Cohen's *d* combined across studies differ significantly from 0? The test for the overall effect, $z = 2.09$, $p = .04$, indicates that the null hypothesis that the overall average effect was zero can be rejected using $\alpha = .05$, two tailed. The mean Cohen's *d* that describes difference of depression scores for CBT compared with control group was –.87. This suggests that average mean depression was almost 1 standard deviation lower for persons who received CBT. That would be labeled a large effect using Cohen's standards (Table 1.1); it lies in between "minimal reportable effect" and a moderate effect using the guidelines of Fritz et al. (2012) (Table 1.2).

Second, are the effect sizes sufficiently similar or close together that they can be viewed as homogeneous? The test for heterogeneity result was $\chi^2 = 2.03$, $df = 2$, $p = .36$. The null hypothesis of homogeneity is not rejected. If the χ^2 test result were significant, this would suggest that some studies yielded different effect sizes than others. If the meta-analysis included numerous studies, it would be possible to look for moderator variables that might predict which studies have larger and which have smaller effects. An actual meta-analyses of CBT effectiveness suggested that effects were larger for studies done in the early years of CBT and smaller in studies done in recent years (Johnsen & Friborg, 2015). In other words, the year when each study was done was a moderator variable; effect sizes were larger, on average, in earlier years than in more recent years.

1.11 RECOMMENDATIONS FOR BETTER RESEARCH AND ANALYSIS

Extensive recommendations have been made for improvements in data analysis and research practices. These could substantially improve understanding of results from individual studies, reduce *p*-hacking, reduce the number of false-positive results, and improve replicability of research results.

Cumming (2012) recommended focusing more on CIs and effect sizes (and less on *p* values) in reports and interpretations of research results. In addition, meta-analyses should be used to summarize effect size information across studies. When effect size information is not examined, small *p* values are sometimes misunderstood as evidence of effects strong enough to be "worthy of notice," in situations where treatment effects may be too small to be valued, and perhaps too small to even be noticed by everyday observation.

Use of language should be precise. It is unfortunate that the phrase "statistically significant" includes a word (*significant*) that means "noteworthy and important" in everyday use. Authors should try to convey accurate information about effect size in a way that distinguishes between statistical and practical significance. If you describe $p < .001$ as "highly significant," this leads many readers to think that the effect of a treatment or intervention is strong enough to be valuable in the real world and worthy of notice. However, *p* values depend on *N*, as well as effect size. A very weak treatment effect can have a very small *p* value if *N* is sufficiently large.

Data analysts need to avoid *p*-hacking, "undisclosed flexibility," and lack of transparency in research reports (Simmons et al., 2011). Authors also need to avoid HARKing: hypothesizing after results are known (Kerr, 1998). HARKing occurs when a researcher makes up an explanation for a result that was not expected. For a detailed *p*-hacking checklist (things to avoid) see Wicherts et al. (2016). When *p*-hacking occurs, reported *p* values can greatly understate the true risk for Type I error, and this often leads data analysts and readers to believe that evidence against the null hypothesis is much stronger than it actually is. This in turn leads to overconfidence about findings and perhaps publication of false-positive results.

The most extensive list of recommendations about changes need to improve replicability of research comes from Asendorpf et al. (2013). All of the following are based on their recommendations. The entire following list is an abbreviated summary of their ideas; see their paper for detailed discussion.

1.11.1 Recommendations for Research Design and Data Analysis

- Use larger sample sizes. Other factors being equal, this increases statistical power and leads to narrower CIs.
- Use reliable measures. When measures have low reliability, correlations between quantitative measures are attenuated (i.e., made smaller), and within-group *SS* terms in ANOVA become larger.
- Use suitable methods of statistical analysis.
- Avoid multiple **underpowered** studies. An underpowered study has too few cases to have adequate statistical power to detect the effect size. Consider error introduced by multiple testing in underpowered studies.

  > The literature is scattered with inconsistent results because underpowered studies produce different sets of significant (or nonsignificant) relations between variables. Even worse, it is polluted by single studies reporting overestimated effect sizes, a problem aggravated by the **confirmation bias in publication** and a tendency to reframe studies post hoc to feature whatever results came out significant. (Asendorpf et al., 2013)

- Do not evaluate whether results of a replication are consistent with the original study by "vote counting" of NHST results (e.g., did both studies have $p < .05$?). Instead note whether the CIs for the studies overlap substantially and whether the sample mean for the original study falls within the CI for the sample mean in the replication study.

1.11.2 Recommendations for Authors

- Increase transparency of reporting (include complete information about sample size decisions, criteria used for statistical significance, all variables that were measured and all groups included, and all analyses that were conducted). Specify how possible sources of bias such as outliers and missing values were evaluated and remedied. If cases, variables, or groups are dropped from final analysis, explain how many were dropped and why.
- Preregister research plans and predictions. For resources in psychology, see "Preregistration of Research Plans" (n.d.).
- Publish materials, data, and details of analysis (e.g., on a webpage or in a repository; see "Recommended Data Repositories," n.d.).

- Publish working papers and engage in online research discussion forums to promote dialog among researchers working on related topics.
- Conduct replications and make it possible for others to conduct replications.
- Distinguish between exploratory and "confirmatory" analyses.

It is obvious that these are difficult for authors to do, particularly those at early stages in their careers. Publication of large numbers of studies that yield statistically significant results is a de facto requirement for getting hired, promoted, tenured, and grant-funded. Publication pressure can lower research quality (Sarawitz, 2016). Requirements to replicate studies and report more detail about data analysis decisions will make the process of publication far more time consuming. Efforts to adhere to these guidelines will almost certainly lead to publishing fewer papers. This could be good for the research field (Nelson, Simmons, & Simonsohn, 2012). Changes in individual researcher behavior can only occur if researchers are taught better practices and if institutions such as departments, universities, and grant-funding agencies provide incentives that encourage researchers to produce smaller numbers of high-quality studies instead of rewarding publication of large numbers of studies.

1.11.3 Recommendations for Journal Editors and Reviewers

- Promote good research practice by encouraging honest reporting of less-than-perfect results.
- Do not insist on "confirmatory" studies; this discourages honest reporting when analyses are exploratory.
- Publish null findings (those with $p > .05$) to minimize publication bias (provided that the studies are well designed). (Of course, a nil result should not be interpreted as evidence that the null hypothesis of no treatment effect is true. It is just a failure to find evidence that is inconsistent with the null hypothesis.)
- Notice when a research report presents an unlikely outcome and raise questions about it. For example, Asendorpf et al. (2013) noted, "If an article reports 10 successful replications . . . each with a power of .60, the probability that all of the studies could have achieved statistical significance is less than 1%," even if the finding is actually "true."
- Allow reviewers to discuss papers with authors.
- Journals may give badges to papers with evidence of adherence to good practice such as study preregistration. *Psychological Science* does this; other journals are beginning to as well.
- Require authors to make raw data available to reviewers and readers.
- Reserve space for publication of replication studies, including failures to replicate.

1.11.4 Recommendations for Teachers of Research Methods and Statistics

To a great extent, textbooks and instructors teach what researchers are doing, and researchers, reviewers, and journal editors do what they have been taught to do. This discourages change. Incorporating issues such as the limitations of p values, the importance of reporting CIs and effect size, the risk for going astray into p-hacking during lengthy data analysis, and so forth, will help future researchers take these issues into account.

- Students need to understand the limitations of information from statistical significance tests and the problems created by inadequate statistical power, running multiple analyses, and selectively reporting only "significant" outcomes. In other words, they need to learn how to avoid *p*-hacking. Some of these ideas might be introduced in early courses; these topics are essential in intermediate and advanced courses. Many technical books cover these issues, but most textbooks do not.
- Graduate courses should focus more on "getting it right" and less on "getting it published."
- Students need to know about a priori power analysis as a tool for deciding sample size (as opposed to the practice of continuing to collect data until $p < .05$ can be obtained, one of many forms of *p*-hacking). Some undergraduate statistics textbooks include an introduction to statistical power. Earlier chapters in this book provided basic information about power for each bivariate statistic.
- The problems with inflated risk for Type I error that are raised by multiple analyses and multiple experiments should be discussed.
- Transparency in reporting should be encouraged. Students need to work on projects that use real data set with the typical problems faced in actual research (such as missing values and outliers). Students should be required to report details about data screening and any remedies applied to data to minimize sources of artifact such as outliers.
- Students can reanalyze raw data from published studies or conduct replication studies as projects in research methods and statistics courses.
- Instructors should promote critical thinking about research designs and research reports.

1.11.5 Recommendations About Institutional Incentives and Norms

- Departments and universities should focus on quality instead of quantity of publications when making hiring, salary, and promotion decisions.
- Grant agencies should insist on replications.

1.12 SUMMARY

The title of an article in *Slate* describes the current situation: "Science Isn't Broken. It's Just a Hell of a Lot Harder Than We Gave It Credit For" (Aschwanden, 2015). Self-correction and quality control mechanisms for science (including peer review and replication) do not work perfectly, but they can be made to work better. Progress in science requires weeding out false-positive results as well as generating new findings. Unfortunately, while generating new findings is incentivized, weeding out false positives is not. *P*-hacking without active intention to deceive is probably the most common reason for false-positive results.

Attempts to identify false-positive results (whether in one's own work or in the work of others) can be painful. Ideally this will happen in a culture of cooperation and constructive commentary, rather than competition and attack. Public abuse of individual researchers whose work cannot be replicated is not a good way to move forward. All of us have (at least on occasion) complained about nasty reviews. We need to remember, when we become upset about the "them" who wrote those nasty reviews, that "them" is "us," and treat one another kindly. Criticism can be provided in constructive ways.

The stakes are high. Press releases of inconsistent or contradictory results in mass media may reduce public respect for, and trust in, science. This is turn may reduce support for research funding and higher education. If researchers make exaggerated claims on the basis of limited evidence, and claims are frequently contradicted, this provides ammunition for antiscience and anti-intellectual elements in our society.

Change in research practices does not have to be all or nothing. It is easy to report CIs and effect sizes (as suggested by Cumming, 2014, and others). Meta-analyses are becoming more common in many fields. We can make more thoughtful assessments of effect sizes and distinguish between statistical and practical or clinical importance (Kirk, 1996; Thompson, 2002a). The many additional recommendations listed in the preceding section may have to be implemented more gradually, as institutional support for change increases.

COMPREHENSION QUESTIONS

1. If Researcher B tries to replicate a statistically significant finding reported by Researcher A, and Researcher B finds a nonsignificant result, does this prove that Researcher A's finding was incorrect? Why or why not?
2. What needs to be considered when comparing an original study by Researcher A and a replication attempt by Researcher B?
3. Is psychology the only discipline in which failures to replicate studies have been reported? (If not, what other disciplines? Your answer might include examples that go beyond those in this chapter.)
4. What does a *p* value tell you about:
   a. Probability that the results of a study will replicate in the future?
   b. Effect size (magnitude of treatment effect)?
   c. Probability that the null hypothesis is correct?
   d. What does a *p* value tell you?
5. "NHST logic involves a double negative." Explain.
6. What does it mean to say that H_0 is always false?
7. In words, what does Cohen's *d* tell you about the magnitude of differences between two sample means? Does *d* have a restricted range? Can it be negative?
8. How does the value of the *t* ratio depend on the values of *d* and *df*?
9. How does the width of a CI depend on the level of confidence, *N*, and *SD*?
10. Review: What is the difference between SE_M and *SD*? Which will be larger?
11. Consider Equation 1.4. Which term provides information about effect size? Which term provides information about sample size?
12. Describe violations of assumptions or rules that can bias values of *p*. Don't worry whether to call something an assumption versus a rule versus an artifact; these concepts overlap.
13. What are the major alternatives that have been suggested to the use of $\alpha < .05$ (NHST)?
14. What is *p*-hacking? What common researcher practices can be described as *p*-hacking? What effect does *p*-hacking have on the believability of research results?
15. What is HARKing, and how can it be misleading?
16. How could *p*-hacking contribute to the problems that sometimes arise when people try to replicate research studies?
17. Is it correct to say that a study with $p < .001$ shows stronger treatment effects than a study that reports $p < .05$? Why or why not?
18. How does theoretical significance differ from practical or clinical significance? What kinds of information is useful in evaluating practical or clinical significance?
19. When people report CIs instead of *p* values, how might this lead them to think about data differently?
20. Can you tell from a graph or bar chart that shows 95% CIs for the means of two groups whether the *t* test that compares group means using $\alpha = .05$ would be statistically significant? Explain your reasoning.

21. If a computer program or research report does not provide effect size information, is there any way for you to figure it out?

22. Explain the difference between ($M_1 - M_2$) and Cohen's *d*. Which is standardized? What kind of information does each of these potentially provide about effect size?

23. In addition to reporting effect size in research reports, discuss two other uses for effect size.

24. What three questions does a meta-analysis usually set out to answer?

25. Find a forest plot (either using a Google image search or by looking at studies in your research area). Unless you already understand odds ratios, make sure that the outcome variable is quantitative (some forest plots provide information about odds ratios; we have not discussed those yet). To the extent that you can, evaluate the following: Does the plot include all the information you would want to have? What does it tell you about the magnitude of effect in each study? The magnitude of effect averaged across all studies?

26. Describe three changes (in the behavior of individual researchers) that could improve future research quality. Describe two changes (in the behavior of institutions) that could help individuals make these changes. Do any of these changes seem easy to you? Which changes do you think are the most difficult (or unlikely)?

27. Has this chapter changed your understanding or thinking about how you will conduct research and analyze data in future? If so, how?

NOTES

[1]The eminent philosopher of science Karl Popper (cited in Meehl, 1978) argued that to advance science, we need to look for evidence that might disconfirm our preferred hypotheses. NHST is not Popperian falsification. Meehl (1978) pointed out that NHST actually does the opposite. It is a search for evidence to disconfirm the null hypothesis (not evidence to disconfirm the research or alternative hypothesis). When we use NHST (with sufficiently large samples), our preferred alternative hypotheses are not in jeopardy. Meehl argued that NHST is not a good way to advance knowledge in the social and behavioral sciences. It does not pose real challenges to our theories and is not well suited to deal with the sheer complexity of research questions in social and behavioral sciences. We make progress not only by generating new hypotheses and findings but also by discarding incorrect ideas and faulty evidence. Selective reporting of small *p* values does not help us discard incorrect ideas.

[2]An exception is that if *N*, the number of data points, becomes very small, the size of a correlation becomes large. If you have only $N = 2$ pairs of *X*, *Y* values, a straight line will fit perfectly, and *r* will equal 1 or –1. For values of *N* close to 2, values of *r* will be inflated because of "overfitting."

[3]Odds ratios or relative risk measures, which can be obtained from logistic regression, are also common effect sizes in meta-analyses. See Chapter 16 later in this book.

DIGITAL RESOURCES

Find **free study tools** to support your learning, including **eFlashcards, data sets, and web resources**, on the accompanying website at **edge.sagepub.com/warner3e**.