2

From Null Hypothesis Significance Testing to Effect Sizes

There are two main parts to the statistical reform argument: the negatives and the positives. The negatives are criticisms of NHST, and the positives refer to advantages of estimation and other recommended techniques. Most of this book concerns the positives, but in this chapter I'll first consider the negatives: NHST and how it's taught and used.

This chapter focuses on

- NHST as it's presented in textbooks and used in practice
- Problems with NHST
- The best ways to think about NHST
- An alternative approach to science and the *estimation language* it uses
- The focus of that language, especially effect sizes (ESs) and estimation of ESs
- Shifting from dichotomous language to estimation language
- How NHST disciplines can become more quantitative

NHST as Presented in Textbooks

Suppose we want to know whether the new treatment for insomnia is better than the old. To use NHST we test the null hypothesis that there's no difference between the two treatments in the population. Many textbooks describe NHST as a series of steps, something like this:

1. Choose a null hypothesis, H_0: $\mu = \mu_0$, where μ (Greek mu) is the mean of the population, which for us is the population of difference scores between the new and old treatments for insomnia. It's most common to choose H_0: $\mu = 0$, and that's what we'll do here.
2. Choose a significance level, most often .05, but perhaps .01 if you wish to be especially cautious.

3. Apply the appropriate statistical test, a *t* test, for example, to your sample data. Calculate a *p* value, where *p* is the probability that, if the null hypothesis were true, you would obtain the observed results, or results that are more extreme—meaning more inconsistent with the null hypothesis.

The *p* value is the probability of obta our observed results, or results that are extreme, if the null hypothesis is true.

4. If $p < .05$ (or whatever significance level you chose), reject the null hypothesis and declare the result "statistically significant"; if not, then don't reject the null hypothesis, and label the result "not statistically significant."

Sometimes, in addition to specifying H_0, an alternative hypothesis, H_1, is also specified. For example, if H_0: $\mu = 0$, then perhaps H_1: $\mu \neq 0$, in which case the alternative hypothesis is two-tailed, meaning we're interested in departures from the null that go in either direction—the new treatment being either worse than or better than the old. When a *p* value is calculated we need to include values that are more extreme than our observed result. Lucky (2008) obtained $M = 3.61$, so the set of results that is "more extreme" or "more favoring the alternative than the null hypothesis" includes values greater than 3.61. However, because the alternative hypothesis is two-tailed, the set must also include results less than –3.61. Calculation of *p* includes results farther from zero than our result, in either direction.

A *Type I error* is the decision to reject H_0 when it's true. The probability of rejecting H_0 when it's true is called the *Type I error rate* and is given the symbol α (Greek alpha). This is the prespecified criterion for *p*, which I referred to previously as "significance level."

The *Type I error rate*, labeled α, is the ability of rejecting the null hypothesis it's true.

A further variation is that the alternative hypothesis may, like the null, be an exact or point hypothesis, H_1: $\mu = \mu_1$. For example, H_1 may be a statement that the new treatment gives, on average, sleep scores 4 units higher on our sleep scale than the old treatment. Specifying such a point alternative allows calculation of *statistical power*, which is the probability of rejecting the null hypothesis if H_1 is true. In other words, if there is a true effect, and it has the exact size μ_1 specified by the alternative hypothesis, power is the probability our experiment will find it to be statistically significant. We'll discuss power in Chapter 12. Specifying a point alternative also allows calculation of the *Type II error rate*, labeled β (Greek beta), which is the probability of failing to obtain a statistically significant result, if H_1 is true. Therefore, power = $1 - \beta$.

Statistical power is the probability of obtaining statistical significance, and thus rejecting the null hypothesis, if the alternative hypothesis is true.

The *Type II error rate*, labeled β, is the ability of not rejecting the null hyp when it is false.

I invite you to examine one or two of the statistics textbooks you are most familiar with, and compare how they present NHST with my description above. Compare the steps in the sequence and the terminology. Note especially what they say about Step 2, the specification in advance of a criterion for statistical significance, perhaps labeled α. Then, for a possibly very interesting comparison, in each book turn to much later chapters where NHST is used in examples. Does the textbook follow its own rules? If its equivalent of Step 2 states that the criterion, or α, must be chosen in advance, in later chapters does it state an α value at the start of each example? Or does it state anywhere that a particular value of α will be used throughout the book? I suspect you might find that, instead, it follows the practice most common in journal articles, which is to calculate and report the p value, then interpret that in relation not to a single prechosen α level, but implicitly in relation to a number of conventional levels, such as .05, .01, and .001. In other words, rather than prespecifying $\alpha = .05$, if you calculate p to be .034, you report the result as statistically significant, $p < .05$, but if you calculate $p = .007$, you claim statistical significance, $p < .01$. If, happily, you obtain $p = .0003$, then you claim statistical significance, $p < .001$. Sometimes use of a set of conventional levels is signaled by asterisks, with more asterisks for smaller p. You could declare the values previously mentioned as statistically significant: .034*, .007**, and .0003***. Sometimes language is used to claim degrees of statistical significance, as when a two- or three-star result is described as "highly statistically significant."

Even if you found the discrepancy I have described between how a textbook introduces NHST and how it uses NHST to analyze data, you may think the difference is no big deal. However, the distinction between (1) setting α in advance and (2) interpreting exact p values, such as .034, on a sliding scale of degrees of statistical significance is vitally important. The two are based on quite different interpretations of the p value. To explain why this matters, I need to describe a little history.

Two Strands in the History of NHST

I'll give here only a very brief sketch of a famous controversy in the early days of NHST. If you are interested in knowing more I recommend Salsburg (2001), which is a book of fascinating stories about famous statisticians and the development of statistics, and which also provides further references. Sir Ronald Fisher made numerous fundamental contributions to statistics, mainly during the first half of the twentieth century. He developed *significance testing*, in which the p value is used as a guide for reaching a judgment about the hypothesis, taking account of all the circumstances. If $p < .01$, he would generally regard the result as clearly significant. He also used .05 as a reference point, although for quite a wide range of p values, perhaps from .01 to .20, he would typically discuss how follow-up experiments could be used to investigate the effect further. Fisher thus regarded p as a measure of strength of evidence against the hypothesis—the smaller the p, the stronger the reason to doubt the hypothesis. He regarded large p values, perhaps $p > .20$, as indicating weak evidence, and he emphasized that such lack of statistical significance should definitely not be taken as meaning the hypothesis is true.

As Salsburg (2001) explains, Jerzy Neyman and Egon Pearson disliked Fisher's approach and developed a more structured form of decision making. They called the hypothesis under test the *null hypothesis* and introduced the *alternative hypothesis*, so their approach became a choice between the two. They required α to be set in advance. The p value was compared with α, and a choice between the null and alternative hypotheses was made according to whether or not $p < \alpha$. Neyman and Pearson also introduced the ideas of *power*, and *Type I* and *Type II errors*.

They considered the Type I error rate, α, as a long-run proportion: If you carry out numerous experiments all with a true null hypothesis, then in the long run, if $\alpha = .05$, you would reject the null hypothesis for just 5% of those experiments. This interpretation of the probability α could only be correct if α were chosen in advance and the data were not permitted to influence the choice of criterion for statistical significance. If $\alpha = .05$ had been chosen, then even $p = .0003$ would lead simply to rejection of the null hypothesis at the $\alpha = .05$ level.

Fisher strongly disagreed with the Neyman–Pearson approach, and both that approach and Fisher's own ideas were extensively criticized. Now, approaching a century later, the criticism continues, but various mixtures of the two approaches are described in numerous textbooks and used by many disciplines as the basis for drawing conclusions from data. Gerd Gigerenzer (1993) described current NHST practices, not as a mixture, but as "an incoherent mishmash" (p. 314) of the ideas of Fisher, and Neyman and Pearson. Raymond Hubbard (2004) referred to an "alphabet soup, blurring the distinctions between p's and α's" (p. 295).

I suspect that many statistics textbooks present NHST using some variation of the steps I set out earlier, which basically take a Neyman–Pearson approach. I also suspect that different disciplines have somewhat different traditions about how NHST is presented. It would be very interesting to know more about how NHST is described in textbooks, and whether that varies over disciplines, but I've been able to find only a few small studies on the topic, and none that make extensive comparisons across disciplines.

It seems the clear structure and decision making of the Neyman–Pearson procedure is appealing. Also, this is the necessary approach if the important topic of statistical power is to be discussed. Therefore, many textbook authors choose it as the framework for presenting NHST. However, researchers seem to have found the requirement to state α in advance too onerous, and unenforceable in practice. It seems irrational to obtain a seemingly clear-cut result with $p = .0003$, but then merely to reject the null hypothesis at the $\alpha = .05$ level. So researchers may teach their students Neyman–Pearson and use that framework to introduce statistical power, but then in practice follow Fisher by reporting exact p values and interpreting these as measures of strength of evidence against the null hypothesis.

If my analysis is even partly accurate, it's not surprising that many students are confused. To some extent students may be expected to learn one rationale and procedure (Neyman–Pearson), but then see a quite different one (Fisher) modeled in the journal articles they read. It would be particularly interesting to investigate whether many textbooks exhibit the discrepancy I described: Do they teach Neyman–Pearson, but then a few chapters later follow Fisher when illustrating how researchers carry out data analysis in practice? It might be tempting to regard a mixture of the two approaches as possibly combining the best of both worlds, but the two frameworks are based on incompatible conceptions of probability. The mixture is indeed incoherent, and so it's not surprising that misconceptions about NHST are so widespread.

I argued in Chapter 1 that NHST can lead to unjustified interpretations of results, whereas estimation provides more complete information. I now turn to a few further problems of NHST and discuss them in terms of how their dangers can be minimized. Box 2.1 reports some relevant evidence from statistical cognition.

BOX 2.1 EVIDENCE OF *p* VALUE MISCONCEPTIONS

In his important book, *Statistical Inference*, Michael Oakes (1986) gave a scathing critique of NHST and a comparison with other approaches to inference. In his Chapter 3 he reported four statistical cognition experiments that explored the statistical intuitions of psychology researchers and postgraduate students. He found evidence of major misunderstanding of NHST, and also misconceptions about other issues we'll discuss in future chapters, including replication and correlation. In his first study he asked six simple true–false questions about the meaning and interpretation of a p value. Only three of his 70 participants answered all six correctly.

Haller and Krauss (2002, tinyurl.com/nhstohdear) presented Oakes's six questions to psychology students and academic staff in six German universities. As in the Oakes study, a simple experiment was described that gave a p value of .01, then the true–false questions were posed. As an example, consider Question 4:

> "You can deduce the probability of the experimental hypothesis being true."

We know that $p = .01$. If this means a 1% probability that the null hypothesis is true, then there's a 99% probability that the null is false and the experimental hypothesis is true. That's just a statement of the common incorrect belief that p is the probability that the results are due to chance. The correct answer to Question 4 is thus "false," but Oakes reported that 66% of his respondents answered "true."

Haller and Krauss (2002) reported that in their sample 59% of psychology students incorrectly answered "true" to that question, as did 33% of psychology academic staff who did not teach statistics. They also obtained responses from psychology academic staff who taught statistics: 33% of these answered "true" to Question 4, and only 20% answered all six questions correctly. Haller and Krauss described that evidence of NHST misconception among teachers of statistics in psychology departments as "flabbergasting" (p. 7). If a technique is not even understood correctly by its teachers, what hope is there for students and researchers who wish to use it?

Statistical cognition research aims to identify problems, but also to find ways to overcome them. Haller and Krauss (2002) did this by discussing how improved teaching might overcome p value misconceptions. Kalinowski, Fidler, and Cumming (2008) reported a small teaching experiment in which they evaluated two approaches: one proposed by Haller and Krauss, and the other an explanation of the basic logic of NHST. Both approaches were reasonably successful in improving students' understanding of p values. However, it remains an enormous challenge to demonstrate that better teaching could overcome all of the many pervasive misconceptions of NHST and, even if we could achieve that, the fundamental problem would remain that NHST focuses only on the narrow question, "Is there an effect?"

There have been other studies of NHST errors, besides those of Oakes (1986) and Haller and Krauss (2002). My conclusion is that evidence of NHST misconception is now strong. Researchers, students, and even teachers of statistics in psychology all have severe and persisting misunderstandings of p values and what they mean. It's hardly surprising that NHST is so often misused.

Some Selected Problems of NHST

The best summary of NHST problems is Chapter 3 of *Beyond Significance Testing* by Rex Kline (2004). The chapter is available from tinyurl.com/klinechap3 as a free download. Kline described 13 wrong beliefs about *p* values and the way they are used and interpreted. He also explained several major ways that reliance on NHST has damaged research and hampered research progress. His chapter concluded with recommendations about how to avoid or minimize NHST problems, and an outline of his version of the new statistics. I'll now discuss some selected problems of NHST, all of which are additional to the central and fundamental limitation that NHST focuses only on "is there an effect?" By contrast, estimation is much more informative.

What *p* Is, and What It's Not

The *p* value is the probability of getting our observed result, or a more extreme result, if the null hypothesis is true. So *p* is defined in relation to a stated null hypothesis, and requires as the basis for calculation that we assume the null is true. It's a common error to think *p* gives the probability that the null is true: That's the *inverse probability fallacy*. Consider the difference between

The *inverse probability fallacy* is the incorrect belief that the *p* value is the probability that the null hypothesis is true.

1. The probability that you speak English if you are reading this book (close to 1, I would think); and
2. The probability that you will read this book, if you speak English. (Even if one million people read this book—I wish!—that's still a probability close to 0, because so many million people in the world speak English.)

Here's another example of the same distinction. Compare

3. The probability of getting certain results if the null is true (that's *p*); and
4. The probability that the null is true if we've obtained certain results. (We'd like to know that but, alas, *p* can't tell us.)

In both these examples, the two probabilities are fundamentally different. Probability 3 is a conditional probability that's easy to calculate if we assume the null is true. Probability 4 refers to truth in the world, and in a sense must be either 0 (the null is false) or 1 (the null is true), but we don't know which. Anyone can choose his or her own subjective probability that the null is true, but different people will most likely choose different values.

Suppose you run a coin-tossing experiment to investigate whether your friend can use the power of her mind to influence whether a coin comes up heads or tails. You take great care to remove any chance of trickery. (Consult a skilled conjurer to discover how difficult that is.) Your friend concentrates deeply then predicts correctly the outcome of nine of the first 10 tosses. A surprising result! You calculate $p = .011$ as the probability that she would get nine or 10 predictions correct, if the null hypothesis of a fair coin and random guessing were true. (That's Probability 3, the p value.) Are you going to declare statistical significance and buy her the drink she bet you? Or will you conclude that most likely she's just had a lucky day? Sure, .011 is small (and less than .05), but you find her claimed power of the mind *very* hard to accept. You need to choose your own subjective probability that she was lucky, and .011 doesn't give you exact help in making your choice. That's the NHST dilemma, which we usually side-step by using conventions that .05 or .01 are reasonable p value cutoffs for declaring a result as statistically significant. We duck the need for subjective judgment by resorting to a mechanical rule that takes no account of the situation.

Note carefully that .011 was the probability of particular extreme results *if the null hypothesis is true.* Surprising results may reasonably lead you to doubt the null, but p is not the probability that the null is true. Some statistics textbooks say that p measures the probability that "the results are due to chance"—in other words, the probability that the null hypothesis is correct. However, that's merely a restatement of the inverse probability fallacy. It's completely wrong to say the p value is the probability that the results are due to chance. Jacob Cohen (1994), a distinguished statistics reformer, wrote that NHST "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (p. 997). In other words, we want to know whether the null is true (Probability 4), but p does not measure that—it measures Probability 3. In desperation we believe that p measures Probability 4 and, unfortunately, some textbooks perpetuate the error.

Beware the Ambiguity of "Significance"

If you read in a journal article that "there was a significant reduction in anxiety," does "significant" mean important or large, or just that $p < .05$ The word is ambiguous and can easily mislead. Any good statistics textbook will explain that statistical significance is different from scientific o practical significance. Small p does not guarantee that an effect is "large or "important." A tiny effect can be highly statistically significant if th experiment is sufficiently large, or a small experiment can find a larg effect that's not statistically significant. Distinguish carefully between interpretation of the effect size (ES), and any NHST statement based on p. Unfortunately, it's fairly common to find a fallacy I call the *slippery slope of significance*: In the Results section of an article, NHST is reported and an effect is declared to be statistically significant because p is small. However, in the Discussion section, and perhaps the abstract, the effect is described simply as "significant" and is discussed as if it's important, or large. The ambiguous term "significant" silently morphs from one of its meanings to the other. Discussing an ES as large or important requires justification based on informed judgment, whether p is large or small.

I refer to the following fallacy as the *slippery slope of significance*: An effect is found to be statistically significant, is described, ambiguously, as "significant," and then later is discussed as if it had thereby been shown to be "important" or "large."

Kline (2004) recommended that we simply drop the word "significant" and write, "there was a statistical reduction in anxiety," if we're reporting NHST and have rejected the null hypothesis. That's a good idea. An acceptable alternative is to say "statistically significant" if that's the intended meaning, and perhaps "practically significant" or "clinically important" if that's your judgment. I try to avoid using "significant" to mean important, and prefer to find some other word. The vital point is that reading the word "significant" should trigger your ambiguity alarm: Does the author make clear what's intended? Is that justified? Beware the fallacy of the slippery slope of significance.

A sidebar like this often gives the definition of a term. If it's a term or expression of my invention I'll usually say so, as I did for the slippery slope of significance, to signal that you probably won't find it in other statistics textbooks.

Beware Accepting the Null

If we conclude that there's a statistically significant advantage of the new treatment for insomnia when there's actually no difference in the population, we're committing a Type I error: We're rejecting the null hypothesis when it's true. NHST limits the risk of Type I errors by requiring small p before you reject the null. On the other hand, if we fail to find statistical significance when there is a population difference, we're making a Type II error: We're failing to reject the null hypothesis, even though it's false. Often, however, little attention is paid to Type II errors. Box 2.2 reports evidence of the low power of published research in many areas of psychology. In practice, in many disciplines, statistical power is often low, meaning that the risk of committing a Type II error is often large. In other words, many experiments have a high chance of failing to detect effects when they do exist. We must therefore be careful not to take statistical nonsignificance (the null is not rejected) as evidence of a zero effect (the null is true).

All good statistics textbooks warn of the danger of accepting a null hypothesis, but the trouble is that the acceptance can be hidden. For example,

BOX 2.2 MANY DECADES OF EVIDENCE OF LOW POWER

Cohen (1962) studied 70 articles in the *Journal of Social and Abnormal Psychology*. He chose ES values to label as "small," "medium," and "large," and then estimated the power of the experiments to find effects of various sizes. He found that estimated power varied greatly over studies, but the mean was only .48 (and median .46) to find medium-sized effects. Mean power to find small effects was .18, and only for large effects was mean power as high as .83. In summary, published research in a number of areas of psychology typically had only about a coin-toss chance of finding a medium-sized population effect to be statistically significant ($p < .05$). Cohen described those levels of power as "far too small" (p. 153) and urged researchers to increase their sample sizes and routinely subject their research plans to power analysis. Cohen noted that almost all the published articles reported statistically significant findings and reasoned that, given the low power, many experiments that failed to find statistically significant effects—thus making Type II errors—must have been conducted, but not published. They represent an enormous waste of research effort. Cohen's study was before the arrival of meta-analysis, but we can now recognize that, because the published studies were a biased subset of all studies conducted, meta-analysis of published studies would give biased estimates—most likely overestimates—of population ESs.

Sedlmeier and Gigerenzer (1989) revisited the same journal 24 years later. They found power to be just as low (median .44 to find a medium-sized effect), and even lower if they took account of alpha adjustment procedures used in many articles to account for multiple tests—procedures not in regular use at the time of Cohen's (1962) study. Sedlmeier and Gigerenzer also reviewed studies of power in about 20 other journals from several disciplines. There was variation but, overall, power was similarly disappointingly low. Later studies have suggested little improvement in the following two decades (Maxwell, 2004), despite the efforts of Cohen and others to persuade researchers to consider power seriously.

in the NHST presentation of Lucky–Noluck the *Inconsistent* interpretati
amounts to concluding that Lucky found an effect but Noluck did not. T
nonrejection of the Noluck null is interpreted as a zero effect. It's absurd
use $M = 2.23$ as evidence that the true value is zero, but hidden under a p
of NHST ritual that's what an *Inconsistent* interpretation does.

I refer to the following fallacy as the *slippery slope of nonsignificance*: An effect is found to be statistically nonsignificant then later discussed as if that showed it to be zero.

If a null hypothesis is not rejected, watch out for another fallacy: the *slippery slope of nonsignificance*. In the Results section, $p > .05$ prompts a statement that the difference failed to reach statistical significance. That's fine, but in the discussion, or even the abstract, the statistically nonsignificant difference may quietly become no difference. Worse, this may be left implicit, as when the result is contrasted with some other, statistically significant effect. The *Inconsistent* interpretation of Lucky–Noluck includes no explicit statement that Noluck found a difference of zero, but making the contrast and stating the two results are inconsistent implicitly assumes that the statistically nonsignificant effect was zero. Example 2.1 is an example. Beware the fallacy of the slippery slope of nonsignificance.

EXAMPLE 2.1 DOES THIS ANTI-AGING CREAM WORK?

In April 2009 queues formed outside some stores in the Boots chain of pharmacies in the United Kingdom as customers rushed to buy No. 7 Protect & Perfect Intense Beauty Serum. Their eagerness was prompted by media reports claiming that an article in the *British Journal of Dermatology* (Watson et al., 2009) provided scientific proof that the product, marketed as an anti-aging cream, actually worked. The original version of the article had just been published online. It was titled "A cosmetic 'anti-ageing' product improves photoaged skin: A double-blind, randomized controlled trial." It stated, "The test product produced statistically significant improvement in facial wrinkles as compared to baseline assessment ($p = .013$), whereas vehicle-treated skin was not significantly improved ($p = .11$)" (p. 420). The article concluded, "An over-the-counter cosmetic 'anti-ageing' product resulted in significant clinical improvement in facial wrinkles" (p. 420).

The article reported a statistically significant improvement for the active ingredient, but no statistically significant improvement for the control treatment, which was "vehicle," meaning cream lacking the ingredient under test. No direct comparison was reported of results for the treated and control participants, and the conclusion was based on the differing *p* values. Do you recognize the pattern as classic Lucky–Noluck? The *p* values (.013 and .11) don't justify the conclusion, and it's a statistical error to claim they do.

Yes, the title of the article sounds convincing, with its statement that the product improves aged skin and that the experiment was a double-blind randomized control trial (RCT). Yes, a double-blind RCT is the gold standard of research designs for this situation. It's not surprising that the media trumpeted the result, but using a good design doesn't guarantee good statistical analysis.

The authors had second thoughts and the official published version, which appeared in the August 2009 printed issue of the journal as well as online, included some amendments. The title became the neutral, "Effects of a cosmetic 'anti-ageing' product on photoaged skin"; it was additionally reported that a comparison of the two conditions gave $p = .10$, and the conclusion claimed only a "[statistically] non-significant trend towards clinical improvement in facial wrinkles." (Watson et al., 2009, p. 419. Note that this is the only version now available. It is the full version originally published online, but with a note explaining the amendments made later by the authors, including changes to the title and the conclusions.) Better science, but perhaps not so likely to trigger media enthusiasm and a rush to buy?

I don't know what prompted the authors' changes, but an article in the journal *Significance* (Bland, 2009) may have contributed. It made the criticism I described previously, as well as other criticisms of the original online version. That original version should not have been published, but we can take this case study as a lesson to never assume that journal referees will find every statistical error. We must always be statistically vigilant and, in particular, watch out for the Lucky–Noluck pattern and the fallacy of the slippery slope of nonsignificance.

If You Use *p*, Report Exact Values

The traditional Neyman–Pearson approach emphasizes making a clear decision to reject the null hypothesis, or not. This requires simply noting whether or not $p < .05$, or, more generally, $p < \alpha$. However, reporting the accurate value of p (e.g., $p = .09$, or $p = .016$) gives extra information and allows any reader either to compare p with a chosen α, or to interpret p as a measure of evidence against the null, as Fisher proposed. The *APA Manual* (2010) states that "when reporting p values, report exact p values (e.g., $p = .031$)" (p. 114), although it permits the use of relative values (e.g., $p < .05$) or asterisks if necessary for clarity in tables. If you use p values, it's best practice and is most informative to report exact values.

If you report a p value, give an exact value ($p = .006$), not a *relative* value ($p < .01$).

How to Think About *p* Values

There has been surprisingly little investigation of how researchers think about p values, the interpretations they make, or their emotional reaction to *p* values that are large or small. These are all great topics for statistical cognition research. One early study was by Rosenthal and Gaito (1963), who asked graduate students and researchers in psychology to rate their degree of confidence in an effect, given a *p* value. They used *p* values ranging from .001 to .9, and found that degree of confidence dropped rapidly as *p* increased. They identified what they called a *cliff effect*, which was a steep drop in the degree of confidence that an effect exists as *p* increased past .05. Poitevineau and Lecoutre (2001) pointed out that Rosenthal and Gaito's cliff effect was steep but only moderate in size, and not shown by all participants. Their own investigation found that different participants showed different patterns of change in confidence as *p* increased from very small to large. One extreme pattern is a smooth decrease in confidence as *p* increases, which fits with Fisher's view of *p* as a measure of strength of evidence. The cliff pattern is quite different, and is a large and sharp drop in confidence at .05, which fits with a Neyman–Pearson dichotomous decision based on $\alpha = .05$. It seems, from the little evidence available so far, that there are elements of each of these patterns in the ratings given by students and researchers, with most showing the gradual decrease and some showing in addition a small or large drop at .05—in other words an element of the cliff effect. There's considerable variation in patterns shown by different respondents, which is consistent with NHST as it is practiced today being a confused mix of Fisherian and Neyman–Pearson ideas.

The *cliff effect* is a sharp drop in the degree of confidence that an effect exists for *p* just below .05 (or another conventional criterion) and just above that criterion.

A good way to think about *p* is in terms of its definition, as a probability of obtaining particular results assuming that the null is true. Whenever you see a *p* value, bring to mind "assuming there's actually no difference," or a similar statement of the null hypothesis. However, I suspect researchers may most commonly follow Fisher and think of *p* as a measure of evidence against the null hypothesis, even though this interpretation of *p* is only occasionally mentioned in statistics textbooks. Other things being equal, the smaller the *p*, the more reason we have to doubt the null. In a later chapter I'll demonstrate that *p* is actually an extremely poor and vague measure of evidence against the null. Even so, thinking of *p* as strength of evidence may be the least bad approach.

There's also a better way to interpret a *p* value: Use it, together with knowledge of the null hypothesis and the obtained mean or other ES, to find the CI, as Chapter 4 explains. Then you'll have a much better basis for interpretation.

I now want to turn from NHST to the new statistics—from the negatives to the positives. I'll start with ESs. I'll introduce ESs by considering some aims of science and the language used to express those aims.

The Questions That Science Asks

Think of the questions that science asks. Some typical questions are

Q1 "What is the age of the Earth?"

Q2 "What is the likely sea-level rise by 2100?"

Q3 "What is the effect of exercise on the risk of heart attack?"

Q4 "What is the relationship between pollution level and fish fertility?"

The answer to Q1 may be 4.54 ± 0.05 billion years, where 4.54 is our best point estimate of the true value, and 0.05 indicates the precision of that estimate. It's natural and informative to answer such quantitative questions by giving a best value and an indication of how accurate you believe this is. Similarly, a news website reports, "Support for the prime minister is 62% in a poll with an error margin of 3%," or you look out from the south rim of the Grand Canyon and say to your friend, "I'm guessing it's 10 kilometers to the other rim, give or take 5 k." The answer to Q2 may be 0.37 m, with a range of 0.18 to 0.59 m. The precision, or uncertainty, is indicated by the error margin, the "give or take," or the range of predictions.

The focus of such questions is an *effect*, and the *size* or nature of that effect. We use *effect* to refer to the age of the Earth, support for the prime minister, or distance to the north rim—in fact anything in which we might be interested. Therefore, an *effect size* (ES) is simply the amount of something that might be of interest. Estimating effect sizes is often the primary purpose of empirical science, and so inevitably the primary outcome is one or more ES estimates. "How much" questions, like Q1, Q2, and Q3, naturally lead to "this much" answers, and the main purpose of journal publication is to report those answers.

An *effect* is anything we might be interested in, and an *effect size* is simply the size of anything that may be of interest.

Many scientists would be astonished to find a chapter in a statistics textbook that explains the importance of reporting ESs. Isn't that obvious? How else could science proceed? NHST disciplines, however, often ask not "how much" questions, but "whether or not" questions, and they publish "statistically significant effect" or "no statistically significant effect" as the dichotomous answer to each.

Chapter 1 described dichotomous thinking, and estimation and meta-analytic thinking. My argument here is that the language used by researchers can indicate which type of thinking predominates. The link may be even stronger: Deliberately choosing estimation language may encourage estimation and meta-analytic thinking, which are the types of thinking favored by the new statistics. Using estimation language to formulate research questions should naturally encourage a focus on ESs and CIs for reporting and interpreting results.

At this point I want to mention the argument of Paul Meehl (1978), the distinguished psychologist and philosopher of science, who published strong criticisms of NHST over several decades. He stated that "reliance on merely refuting the null hypothesis … is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology" (p. 817). Not only was dichotomous decision making impoverished, he argued, but it limited the research aims, and even the theories that researchers formulated. He blamed "the Fisherian tradition [NHST], … [which] has inhibited our search for stronger tests, so we have thrown in the sponge and abandoned hope of concocting substantive theories that will generate stronger consequences than merely 'the Xs differ from the Ys'" (p. 824). He urged researchers to abandon NHST, and to build theories that were more quantitative, so they could "generate numerical point predictions (the ideal case found in the exact sciences)" (p. 824). Meehl is saying that, for example, focusing on how much better the new Lucky–Noluck treatment for insomnia is than the old should encourage us to consider a quantitative model of how the treatment works. More generally, the new statistics could lead to better theory as well as more informative research, so disciplines could become more quantitative and thus more sophisticated, and better able to explain the world.

The Language Researchers Use

I would like to be able to report evidence that shifting to estimation language encourages the use of ESs and CIs, and leads to better research and more informative results. Lacking such evidence, I would at least like to be able to present evidence about how often researchers use dichotomous or estimation language to present their research aims and report their conclusions. Box 2.3 reports the only study I know of on these issues, and it presents only a partial picture: It didn't examine how research aims are expressed, but found evidence in a leading psychology journal that conclusions are often expressed in dichotomous language.

As an initial tiny investigation of researchers' language I picked up the most recent issue of *Psychological Science*, a leading journal that reports interesting findings across the whole of psychology. I scanned the first 10 articles, looking for brief statements of an article's main aim or question. (I'm not going to give referencing details for that issue, or the words I quote from those articles, because I'm presenting them as typical generic

BOX 2.3 WHAT LANGUAGE DO RESEARCHERS USE?

Hoekstra, Finch, Kiers, and Johnson (2006) examined 266 articles published in *Psychonomic Bulletin & Review*, a leading psychology journal, during 2002 to 2004. They found that 97% of the articles used NHST, and only 6% reported any CIs. They also found that 60% of the articles that used NHST reported a statistically nonsignificant result then made the serious mistake of accepting the null hypothesis and claiming no effect—which sounds like the fallacy of the slippery slope of nonsignificance. In addition, 19% of the articles that used NHST reported a statistically significant effect, then those articles made the mistake of stating, on the basis of statistical significance, that the effect certainly existed, or was important. That sounds like the fallacy of the slippery slope of significance. Hoekstra et al. (2006) interpreted these mistakes as evidence that a majority of researchers did not appreciate the uncertainty in any NHST result, and stated their conclusions using dichotomous language that implied certainty.

fragments of language and don't want to imply criticism of particular authors when I'm quoting only very few of their words.)

I must say I was impressed. Every article investigated interesting issues and reported ingenious experiments that provided relevant evidence and led to valuable conclusions. *Psychological Science* accepts only 10–20% of the manuscripts it receives, and the published articles describe some of the best psychological research from anywhere around the world. However, all 10 articles used NHST to identify effects as statistically significant or not, so dichotomous thinking still thrives. In two of the 10, significance language was avoided, but, even in these two articles, effects were still described as existing or not: p values provided the basis for statements like "proud participants … spent more time manipulating the puzzle … $p = .04$ …," and "no difference was found between reaction times for the proximal and distal postures … $p > .05$…." The reader is left to insert "statistically significant(ly)" at the appropriate point in each sentence. One example diverged from the usual $p < .05$ criterion for regarding an effect as real: "They were also perceived as more dominant by their partners … $p = .07$." A null hypothesis was accepted with the statement, "As Figure 3 shows … the magnitude of this effect was indistinguishable from the magnitude of the adaptation with consistent illumination … $t(13) = 1.82$, n.s." The two points referred to were an easily distinguishable 4 mm apart in Figure 3, which is hardly good evidence for a zero difference, especially given that the unstated p value was .09.

All 10 articles reported means or other ESs, often in figures. Discussion often referred to the ESs. Aims were often expressed in a general way:

"The goal of this study was to explore the mechanism by which gesturing plays a role in learning." Even so, in eight articles it was easy to find a main aim and corresponding conclusion each expressed in a dichotomous way. Here are some examples:

> *Aim 1*: "We predicted that playing a violent video game … would decrease the likelihood of help."
>
> *Conclusion 1*: "Participants who played a violent game took significantly longer to help."
>
> *Aim 2*: "We hypothesized that stressed participants would exhibit increased risky behavior on loss-domain trials but increased conservatism on gain-domain trials."
>
> *Conclusion 2a*: "Significantly fewer risky decisions (i.e., increased conservatism) were made on gain-domain trials under acute stress."

The usual $p = .05$ cutoff was relaxed in order to make the following conclusion related to the first part of the aim:

> *Conclusion 2b*: "On loss-domain trials, participants showed a trend toward making a higher number of risky decisions under acute stress … $p < .10$."

Just two studies used estimation language and included no dichotomous statements of aims. One stated, "The current study measured the degree to which the public's interpretation of the forecasts … matches the authors' intentions." "Measured the degree to which" is the crucial estimation language. Discussion focused on the extent of differences. The aim of the other study was "estimating the financial value of pain." Here "estimating the value of" is the crucial wording. Discussion focused on the price people would pay to avoid pain in various circumstances. Both studies used significance language as well, but were infused with estimation language and estimation thinking, and focused on ES estimates as answers to the research questions asked. Just one of the 10 articles reported CIs, shown in a figure. However, the CIs were not mentioned in the text nor used for interpretation.

I conclude that much excellent current research relies on NHST, dichotomous thinking, and significance language, while also reporting ESs. I can't be sure, but I suspect that at least some of the eight articles that used dichotomous language could have been more informative if estimation ideas had shaped the questions and guided the data analysis and interpretation. The two studies that used estimation language to express their aims, and focused in their discussions on estimated ESs, suggest how future research might be different. Choosing estimation language and CIs, and keeping the focus on ESs, should avoid errors caused by NHST, while also giving answers that are more quantitative. This should contribute to the broader goal of building more quantitative theories, and more quantitative and progressive disciplines.

I suggested previously that scientists in disciplines that don't rely on NHST automatically assume that ESs are the focus of most experiments. I suspect that researchers in NHST disciplines share this intuition about wanting a numerical value from an experiment, and information about how good an answer to our question that value is. The problem is that NHST and dichotomous thinking to some extent suppress those intuitions. If I'm right, moving to estimation thinking and the new statistics will allow such intuitions to flourish, and researchers may thus feel the new statistics are somewhat natural, or even familiar.

Effect Sizes

Jacob Cohen (1990) wrote, "The primary product of a research inquiry is one or more measures of effect size.... Effect-size measures include mean differences (raw or standardized), correlations ... whatever conveys the magnitude of the phenomenon of interest appropriate to the research context" (p. 1310). This accords with my previous definition of an ES as the amount of anything that might be of interest. It can be as familiar as a mean, a difference between means, a percentage, a median, or a correlation. It may be a standardized value, such as Cohen's *d* (more on this later), or a regression coefficient, path coefficient, odds ratio, or percentage of variance explained (don't worry if some of these aren't familiar). The answer to Q3, about the effect of exercise on risk of heart attack, may be expressed as a percentage change in risk, or a decrease in the number of people in 1,000 who are likely to have a heart attack in one year, or in various other ways, as I will discuss in Chapter 14, but in any case the answer is an ES.

If you measure the attitudes of a group of people before and after you present them with an advertising message, it's natural to think of the change in attitude as an *effect* and the amount of change as the *size* of that effect. However, the term "effect size" is used much more broadly. It can refer to an amount, rather than a change, and there need not be any easily identifiable "cause" for the "effect." The mean systolic blood pressure of a group of children, the number of companies that failed to submit tax returns by the due date, and the ratio of good to bad cholesterol in a diet are all

A *p* value is *not* an ES. It cannot provide the answer to an estimation question.

TABLE 2.1

Examples of ES Measures

Sample ES	Description	Example
Mean, *M*	Original units	Mean response time, $M = 462$ ms.
Difference between two means	Original units	The average price of milk increased last year by \$0.12/L, from \$1.14/L to \$1.26/L.
Median, *Mdn*	Original units	Median response time, $Mdn = 385$ ms.
Percentage	Units-free	35.5% of respondents were in favor. 0.7% of responses were errors.
Frequency	Units-free	39 states ran a deficit.
Correlation, *r*	Units-free	Income correlated with age ($r = .28$).
Cohen's *d*	Standardized	The average effect of psychotherapy was $d = 0.68$ (see Chapters 7 and 11).
Regression weight, *b*	Original units	The slope of the regression line for income against age was $b =$ \$1,350/year.
Regression weight, β	Standardized	The β-weight for age in the regression was .23.
Proportion of variance, R^2	Units-free	Three variables of age, education, and family status in the multiple regression together gave $R^2 = .48$.
Risk	Units-free	The risk that a child has a bicycle accident in the next year is 1/45.
Relative risk	Units-free	A boy is 1.4 times as likely as a girl to have a bicycle accident in the next year.
Proportion of variance, ω^2 (Greek omega-squared)	Units-free	In the analysis of variance, the independent variable age accounted for $\omega^2 = 21.5\%$ of total variance.

perfectly good ESs. Yes, many things are ESs, but *p* values are not. Let me emphasize that last point: A *p* value is *not* an ES, and it cannot provide the answer to a "how much?" estimation question.

Table 2.1 presents some example ESs. Don't worry if some are unfamiliar. In later chapters we'll discuss some of the ESs in the table further, especially Cohen's *d*, Pearson's correlation *r*, and proportions. We'll also encounter further ESs. One of the messages of Table 2.1 is that there are many different ESs that can be described and classified in various ways. Another is that many ESs are probably very familiar to you already. Some authors write about ESs as if they are complex and unfamiliar, and the requirement to report ESs is a new and challenging demand for researchers. Indeed, some ESs may be unfamiliar and a little tricky to understand, but many ESs are highly familiar, and many researchers have always reported and discussed in their articles at least basic ESs, such as means or correlations.

All through the previous paragraph I wrote "ESs," although "ES measures" may have been more precise wording. Many writers use "effect size" to refer sometimes to a single value, and at other times to the measure, such as *M* or *d*, as I did previously. Often I'll write "ES measure" for clarity, but you need to be aware that "ES" is used in these two ways.

The *population ES* is simply the true value of an effect in the underlying population. The *sample ES* is calculated from the data and is typically used as our best point estimate of the population ES. It is often referred to as an *ES estimate*. Estimated ESs are usually the main results of research, and should be the main focus of interpretation because they are the best information we have about the population.

We calculate from our data the *sample ES* and use this as our estimate of the *population ES*, which is typically what we would like to know.

What the *Publication Manual* Says

The *Publication Manual* (APA, 2010) is clear about the necessity of reporting ESs:

> For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size.... Effect sizes may be expressed in the original units (e.g., the mean number of questions answered correctly; kg/month for a regression slope) and are often most easily understood when reported in original units. It can often be valuable to report an effect size not only in original units but also in some standardized or units-free unit (e.g., as a Cohen's *d* value) or a standardized regression weight. (p. 34)

Sometimes a result is best reported both in original units, for ease of understanding by readers, and in some standardized measure for ease of inclusion in future meta-analyses. There will be examples in later chapters. Sources of advanced advice about ESs include Kirk (2003) and Grissom and Kim (2005).

Further important advice from the *Publication Manual* is the requirement to "mention all relevant results, including those that run counter to expectation; be sure to include small effect sizes (or statistically nonsignificant findings)" (APA, 2010, p. 32). If, as often in the past, a finding is only reported in detail, with its ES, if it reaches statistical significance, then the published literature contains a biased sample of research findings. Other things being equal, smaller ES estimates are less likely to reach statistical significance and so would be more likely to remain unpublished, and thus be at risk of being omitted from meta-analyses. If that happens, meta-analysis of published research would be likely to give overestimates of population effects. It's an important part of meta-analytic thinking to understand that *any* ES found by any well-conducted experiment needs to be available for later meta-analysis, whether the ES is small or large, statistically nonsignificant or significant.

Cohen's Reference Values

Jacob Cohen, the statistical reformer I've mentioned a couple of times, championed the use of statistical power. He hoped that if researchers routinely calculated and reported power, they would realize that power is often very low and so may be prompted to design larger and better experiments with higher power. Box 2.2 describes his early work, which revealed the very low power of much published research. Cohen's (1988) *Statistical Power Analysis for the Behavioral Sciences* is his classic book on statistical power that remains a basic reference. Power, as I discussed earlier in this chapter, requires specification of an exact population ES. Cohen therefore needed ES measures for many situations, and *d* is his basic standardized ES, and the topic of Chapter 11. Now, *d* is expressed in SD units, and is thus a kind of *z* score. Consider, for example, IQ scores, which are often expressed on a scale that has SD = 15 in a large reference population. A difference of 7.5 IQ points is half of one SD, or equivalently $d = 0.50$. A sample value of *d* can be used to estimate the corresponding population ES of Cohen's δ (Greek delta).

Cohen urged researchers to interpret their ESs by making an informed judgment in the research context. He also suggested values that might be regarded as "small," "medium," and "large." For *d* these were 0.2, 0.5, and 0.8, respectively. Therefore, a difference of 7.5 IQ points could be regarded, according to Cohen's suggestion, as a medium-sized effect. Cohen favored knowledgeable interpretation in the situation, but offered his values as a "conventional frame of reference which is recommended for use only when no better basis ... is available" (1988, p. 25). His values were, however, shrewdly chosen, and are reasonable for use in some, but far from all, situations. Cohen also suggested .1, .3, and .5 as small, medium and large values of Pearson's correlation *r*.

Interpreting ESs

A focus on ESs can change the way results are reported and discussed. NHST might prompt a researcher to report "children were significantly less anxious after hearing the music, $t(23) = 2.50$, $p = .02$ (two-tailed)" and conclude that "the music significantly lowered children's anxiety." (When you read those statements, did you automatically insert "statistically" before each "significantly"?) That's dichotomous thinking in action, and it tells us little about the experimental result. The mean anxiety scores may, or may not, be reported, but the focus is on *p* values and statistical significance. It would be much more informative to report that "after hearing the music, the decrease in children's anxiety scores had mean $M = 5.1$ units on the anxiety scale, 95% CI [0.88, 9.32]. A decrease of 5 on the anxiety scale is practically beneficial." The focus is on the size of the difference, and the

CI tells us about the precision of the estimated ES. The "practically beneficial" comment is an interpretive judgment by the researcher, based on knowledge of children, the anxiety scale, and the full context. It should be accompanied by a justification. Another researcher may prefer a different interpretation of the ES, but the CI provides information that assists any reader to assess an author's interpretation. The primary interpretation of research should be such judgments about ESs, rather than ritualistic statements about statistical significance based on p values. (Did you notice that the 95% CI does not include zero? You can thus declare the result statistically significant, $p < .05$, if you wish.)

Look back again at those NHST and CI results. Do you feel uncomfortable? Is the conclusion "the music [statistically] significantly lowered ..." still appealing because it seems so clear-cut, so reassuring? In comparison, the very wide CI may be unsettling, even unbelievable. A large amount of careful work over several months, and all we can say is that improvement in anxiety was, most likely, somewhere between about 1 and 9? Sorry, but [0.88, 9.32] corresponds to $p = .02$. (More on that in Chapter 4.) The CI message is accurate, and the apparent certainty of NHST is misleading. We need to come to terms with the large uncertainty in most experimental results and not blame the CIs. Don't shoot the messenger. Appreciating the extent of uncertainty should lead, as I argued in Chapter 1, to meta-analytic thinking and a search for opportunities to cumulate evidence over experiments.

Alternatively, you may be thinking that a statement of significance is more objective—simply note whether or not $p < .05$—whereas the "practically beneficial" interpretation is mere opinion. Yes, interpretation of ESs, like numerous other aspects of research, requires judgment, but readers can make their own interpretations of the published ES values if they wish. I will discuss interpretation of various types of ESs in later chapters. Most basically, the focus should be on ESs because they are of greatest interest, and estimates of ESs are highly informative. The research investigated the effect of music on anxiety, and what we most want to know—what is most valuable for a music therapist—is how large a reduction in anxiety music may give. That's simply the ES, and the CI indicates how good an estimate it is. ESs and CIs together should usually provide the best basis for understanding results.

Sometimes a researcher is fortunate and can choose ESs likely to be familiar to all readers. Examples include height in meters, outcome in number of deaths, value in dollars, and temperature in degrees Celsius. Almost as fortunate is the opportunity to use ESs that are very familiar in the discipline. Thus particular disciplines routinely use response time in milliseconds, attitudes on a 1-to-7 Likert scale from strongly disagree to strongly agree, ability expressed as an age-equivalent score, or blood pressure in millimeters of mercury. Researchers can report and discuss these types of ESs with little ceremony, confident that their readers will have a basic understanding. Even so, they will probably need to explain what particular values or differences mean in the research context, and give reasons to justify interpretive judgments that an ES is, in the context, "large" or "important."

Table 2.1 includes units-free ESs (e.g., Pearson correlations on a scale from –1 to +1) and standardized ESs (e.g., Cohen's *d*, a number of SDs). In many cases a researcher can assume that these will be familiar to readers, although, again, particular values may need explanation in the context, whether or not the researcher chooses to use Cohen's reference values for small, medium, and large. Other ESs are less well known, for example, ω^2 (Greek omega-squared) as a measure of the proportion of variance attributable to an independent variable in an analysis of variance. They may need further explanation, although often a research field develops traditions of using particular measures, and so even a generally little-known ES may be familiar to researchers in that field. In such cases the field may also develop its own reference standards for what's regarded as large or small, important or trivial. Researchers should be wary, however, of writing for only their close colleagues, and should consider as broad a target audience as possible. Research results need to be widely available and reported so practitioners, for example, can readily understand.

I've mentioned Cohen's reference values for some ESs. Many of the tests used in education, psychology, and other disciplines include reference values that can similarly be used to assist interpretation. The Beck Depression Inventory is an example: For the BDI-II (Beck, Steer, Ball, & Ranieri, 1996), scores of 0–13, 14–19, 20–28, and 29–63 are labeled, respectively, minimal, mild, moderate, and severe levels of depression. As a less formal example, a neuropsychologist colleague of mine describes a rough guideline he uses: A decrease of about 15% in the memory score of a client with some brain injury, between two testing times, is the smallest difference he judges likely to be clinically noteworthy. It would be great if increased attention to ES interpretation encourages researchers to develop further formal or informal conventions for what various sizes of effect mean in various contexts.

My conclusion is that a researcher first needs to judge how much can be assumed and how much needs explanation about an ES measure itself, and then should explain and interpret the particular values being reported. Examples 2.2 illustrate a wide variety of types of ESs and various strategies researchers can use to explain and justify their interpretations.

EXAMPLES 2.2 REPORTING AND INTERPRETING EFFECT SIZES

Volcanic Residues in Lake Sediments

Schiff et al. (2008) reported a study of a core sample taken from a lake bed in Alaska. They found 67 layers of tephra—volcanic ash—that had been deposited by 67 eruptions of a nearby volcano. They were interested in the timing of those eruptions, which they estimated from the depths of tephra layers in the core. They carried out several types of mathematical modeling, with much use of CIs. That was all rather complicated, but the main ESs were simple: The researchers discussed depths, layer thicknesses, and particle sizes, all expressed in centimeters; and times in years. They provided graphics to illustrate how depth in the core (expressed in cm) translated into years into the past, and gave tables reporting the size of the ash particles in the various layers and the thickness of those layers. The layers ranged in thickness from 0.1 cm to 8.0 cm, and depths from 3 cm to 562 cm. Estimated eruption times ranged from the present back to 8,660 years ago. The reporting is so comprehensive, and the ESs measures (cm and year) so familiar, that any reader can understand the main findings. The researchers were fortunate in being able to use such familiar ESs, but they took full advantage by presenting well-designed graphs and a clear discussion of the numerous values they observed.

Portion Sizes and Children's Eating

Fisher and Kral (2008) investigated how portion sizes of presented food influenced how much children chose to eat. They discussed portion size in grams and amount eaten in grams and kilocalories. They used percentages freely. The ESs are sufficiently familiar for readers to easily understand the discussion. They spoke, for example, of their adolescents drinking "75% more juice when using a short ..., wide glass than a taller ..., narrower glass of the same volume" (p. 43), and stated that "Doubling the ... size of ... a snack ... increased energy intake ... by 22% (~180 kcal)" (p. 41). The authors also defined the energy density of food; this measure may be unfamiliar to readers, but the explanation given and its close link to weight in grams and energy in kilocalories mean that readers can understand. A typical conclusion was, "When the ED [energy density] of an entrée was

... reduced by 30%, 2- to 5-year-olds consumed 25% fewer [kilo]calories" (p. 41). Fisher and Kral used ES measures that were either very familiar or based on familiar measures, and therefore they could expect readers to readily understand their results and conclusions.

Dropping Out of Medical School

Dyrbye et al. (2010) surveyed students in a number of medical schools to study burnout and serious thoughts of dropping out. Their major ESs were the survey scores, so they described the questionnaires in the survey and referred to previous research that provided evidence of reliability and validity of the measures. They described in detail their own simple scale of "seriousness of thoughts of dropping out," which ranged from "not seriously" to "extremely seriously." They gave reference values for some measures, for example, by reporting that "mental quality of life" (QOL) scores have a mean of 49.2 and SD of 9.5 for the whole U.S. population. They also reported "pre-established thresholds for health professionals" for several of the measures—for example, any score below 33 on the "low sense of personal accomplishment" scale is regarded as "low" for health professionals. These full descriptions and reference values gave a good basis for their discussion of the scores for their medical students and interpretation of relationships between burnout measures and thoughts of dropping out. Once we know about the QOL measures, for example, we can grasp a summary statement that, other things being equal, a one-point-lower QOL score means a student is on average 5% more likely to have serious thoughts of dropping out during the following year. The main ES measures are unlikely to be familiar to readers, so full descriptions and reference values are needed, but then we can understand.

Improvement in Reading Ability

Edmonds et al. (2009) reported a meta-analysis of intervention studies that sought to improve the reading of teenagers with reading difficulties. They used "effect size, *d*" (p. 266) as their main measure, explained how it was calculated, then used Cohen's reference values (0.2, 0.5, and 0.8) in their discussion. McGuinness (2004) also used *d* as the main ES in her large and impressive review of research on reading. (I discuss her work further in Chapter 7.) She didn't mention Cohen's reference values, but made statements such as, "Effect sizes were low for comprehension (.10), marginal for word recognition, spelling, and writing (range .30 to .34), moderate for phoneme awareness (.56), and large for nonword decoding (.71)" (p. 148). She described a value of 0.74 as "solid" (p. 127). Her interpretations were thus largely consistent with Cohen's reference values.

These two examples illustrate how a standardized ES, Cohen's *d*, has become widely used and can serve as a good basis for discussion and interpretation. Researchers can use Cohen's reference values or make their own evaluations of size, but should justify their choice in the particular context.

Risk of Colon Cancer

Moore et al. (2004) studied the relation between obesity and the risk of colon cancer. I'll use their study to discuss two ESs: body mass index (BMI) and risk. BMI is calculated as a person's weight in kilograms divided by the square of his or her height in meters. Some experts criticize BMI because it makes no distinction between fat and muscle, but here I'll follow Moore et al., who used BMI and the World Health Organization definitions of BMI <25 as normal, BMI between 25 and 30 as overweight, and BMI >30 as obese. Such reference values are useful for interpretation, but may change as research advances over the years, and different authorities may recommend different cutoffs. The risk of colon cancer is estimated as the proportion of people in a particular group who developed the cancer during a defined period. Relative risk is the ratio of the risks for two different groups of people. One conclusion was that, for people aged 30 to 54, other things being equal, being obese rather than of normal weight is associated with an increase in risk of colon cancer from 1.2 to 1.8 in 1,000, which is a 50% increase in the risk. That's reasonably understandable, but note that it's important to be told the risks as well as the percentage increase, because a 50% increase in risk may have different implications if it's an increase from a one-in-a-million to a 1.5-in-a-million risk, or an increase from a 10% to a 15% risk. Each of those is a 50% increase in risk. (There's more on this in Chapter 14.) Overall, Moore et al. gave sufficient explanation of BMI and risk, and of what various values mean, for their results and conclusions to be understandable to most readers.

Greasing the Wheels

My final example is a cautionary tale. The title of Kim and Ruge-Murcia's (2009) article asks, "How much inflation is necessary to grease the wheels?" The researchers developed a complex mathematical model of one version of an ideal economy. Many strong assumptions were required. They then applied the model to the U.S. economy and concluded that an inflation rate of 0.35% per year leads to optimum results—if their model is correct and all the assumptions apply. Their ES of 0.35% for the inflation rate is familiar and easily grasped by most readers. We might possibly regard it as a low and desirable rate of inflation. However, figuring out what the whole study means in practice, and the extent to which the model and all the assumptions are realistic, is not nearly so straightforward. The simple and familiar final ES may even be misleading, if it tempts readers to overlook the complex underlying theory and its stringent assumptions. No ES can be better than the data and models on which it's based.

The main new statistics message of this chapter is that the focus of research should almost always be ESs. Adopt estimation thinking, use estimation language to express research goals as "how much" questions about ESs, report ES estimates with their CIs, then interpret those estimates. However, there are also new statistics goals beyond ESs and CIs. For example, the answer to Q4 may be a negative correlation between pollution level and fish fertility. A correlation is a perfectly fine ES, but even better may be a function that expresses the relation between the variables. The journal article may present a figure that plots how fertility changes as a function of pollution; there may even be an equation that describes the relation. Beyond "how much" questions are "what is the relation between" questions, whose answers can be even more informative, and the basis for what Paul Meehl (1978) wanted: theories and disciplines that are more quantitative. I hope the new statistics I discuss in this book encourage researchers to go further and develop and test such quantitative models.

It's time for take-home messages. I invite you to write your own before looking ahead to mine. The preceding paragraph includes some reminders.

Exercises

2.1 Choose a journal of interest to you that reports empirical research. It's best if it has relatively short articles. Find examples of poor NHST, including ambiguous use of the word "significant," and use of relative p (e.g., $p < .01$) rather than exact p (e.g., $p = .006$).

2.2 Find an example of the fallacy of the slippery slope of significance: In the Results section of an article an effect is declared "significant," or even "statistically significant," because p is small. In the discussion, or the abstract, that effect is referred to as certain, or notable, or large, without any attempt to justify such an interpretation of the ES.

2.3 Find examples of acceptance of a null hypothesis. Give yourself extra points for finding an example of the fallacy of the slippery slope of nonsignificance: An innocuous statement of "statistical nonsignificance" in the Results section becomes a statement of no difference in the discussion or the abstract.

2.4 Read as much of Chapter 3 of Kline (2004, tinyurl.com/klinechap3) as you find interesting. Focus on pp. 61–70 and 85–91. Decide what attitudes you'll take to NHST that you read in journal articles.

2.5 Read as much of Haller and Krauss (2002, tinyurl.com/nhstohdear) as you find interesting. Do the findings strike you as astonishing, or depressing? Print out the short questionnaire (in Section 2.2 of Haller and Krauss) including the six Oakes questions. Try them out on your friends—and maybe your teachers?

2.6 Scan a few articles that report experimental results. In each article, first identify a concise statement of the main experimental aim, if possible in relation to a single variable. Can you find a dichotomous statement of that aim? Is there also a "how much" statement of the same aim?

2.7 Match some answers to the corresponding aims. Can you find a dichotomous statement that answers a dichotomous question? Can you find a "this much" answer to a "how much" question? In each case, go further and examine the conclusion or interpretation: Is it expressed in dichotomous or estimation language?

2.8 Find a few dichotomous statements of hypotheses or experimental aims. For each write down your own corresponding "how much" question.

2.9 Here's a challenging one: Find two results in a single article that are compared and that have the pattern of Figure 1.1. As in Lucky–Noluck, one should be statistically significant and the other not. (They need not appear in the article's introduction, but could be results reported for two different groups or conditions. See Example 2.1.) How are they compared? Is it concluded that they are different? Is there discussion about why they might be different? Is a null hypothesis implicitly accepted? If so, what wording is used to hide that? Suggest a better way to report and interpret the two results.

2.10 Look back at Figure 1.1. I criticized the strategy of concluding *Inconsistent* just because one result is statistically significant and the other is not. Could you use NHST to examine whether *Inconsistent* is justified? If so, how?

Exercises 2.9 and 2.10 are based on the previous chapter. There's evidence that it helps learning to include questions about issues discussed earlier, so it's best not to just skip them. Also try to find links with the current chapter.

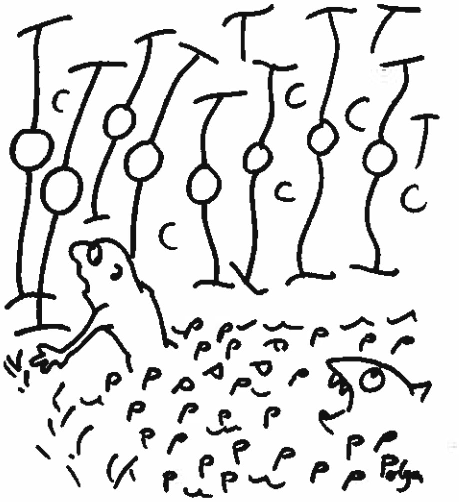
2.11 Work at becoming a "new statistics aware" reader, always alert to misinterpretations caused by NHST, always asking the corresponding "how much" question.

2.12 Find examples of different types of ES estimates. They may be in the text, in tables, or in figures. Look for a mean, a percentage, a measure of change, and a correlation. For any statistical technique that you know about (analysis of variance, regression, chi-square, path analysis, factor analysis, etc.), look for examples of ES estimates.

2.13 For some of your ES examples, find where the ES is discussed or interpreted. Can you find at least one where the ES is given a substantive interpretation? In other words, find a statement about the meaning or importance of the observed size of the effect.

2.14 For an ES without such an interpretation, try to offer your own substantive interpretation: In your judgment, how important, or meaningful, or practically useful is an effect of the size observed? Justify your answer.

2.15 Revisit your take-home messages. Improve them and extend the list if you can.



Take-Home Messages

- NHST as practiced in many disciplines is an uneasy mixture of Fisher's idea that p is a measure of strength of evidence and the strict Neyman–Pearson rule to choose α in advance then decide between null and alternative hypotheses according to whether or not $p < \alpha$.
- Beware NHST traps. A p value is a tricky conditional probability, assuming that the null hypothesis is true. It is not the probability that the results are due to chance.
- Whenever you read a p value, automatically think, "assuming the null hypothesis is true."
- If reporting a p value, give an exact value, not merely a statement like $p < .05$.
- Beware the ambiguous word "significant." Use it with great care, or avoid it.
- Statistical significance is different from practical importance—as you probably knew. But keep the distinction carefully in mind anyway. Beware the fallacy of the slippery slope of significance.

- Avoid accepting a null hypothesis, even implicitly. Beware the fallacy of the slippery slope of nonsignificance.
- Regarding *p* as a very rough index of strength of evidence against the null may be the least bad way to think about *p*.
- Notice the language used to express research aims and conclusions. Wherever possible, prefer estimation language ("how much ...?," "to what extent ...?") to dichotomous language ("is there a difference ...?").
- An effect size (ES) is simply an amount of something that might be of interest. ES estimates from data are our best guide to population ESs. ESs can be as familiar as a difference between means, a percentage change, or a correlation.
- The focus of research is usually effects. Report ESs and wherever possible the CIs, too.
- Interpret ES estimates, using knowledge of the research area and judgment, and justify the interpretation. Cohen's conventional values may be useful. To what extent is each ES large or small, important or unimportant, useful to practitioners?
- The aim is to use estimation language, and estimation, in order to build more quantitative disciplines that make better research progress.