1

Introduction to The New Statistics

This book is about how to picture and think about experimental results. I'll start with a simple pattern of results you may have seen many times, but first I should say what this chapter is about. It introduces

- Null hypothesis significance testing (NHST) and confidence intervals (CIs) as two different ways to present research results
- Meta-analysis as a way to combine results, and thus a third way to present them
- The desirability of shifting emphasis from NHST to CIs and meta-analysis, which are important parts of what I'm calling *the new statistics*
- Three ways of thinking that correspond to the three ways to present results
- Evidence-based practice in statistics, statistical cognition, and some relevant evidence
- ESCI (pronounced "ESS-key," Exploratory Software for Confidence Intervals)

A Familiar Situation: Lucky–Noluck

Consider first a simple pattern of results you may be familiar with.

First Presentation: NHST

Suppose you read the following in the introduction to a journal article:

> Only two studies have evaluated the therapeutic effectiveness of a new treatment for insomnia. Lucky (2008) used two independent groups each of size $N = 22$, and Noluck (2008) used two groups each with $N = 18$. Each study reported the difference between the means for the new treatment and the current treatment.
>
> Lucky (2008) found that the new treatment showed a statistically significant advantage over the current treatment: $M(difference) = 3.61$, $SD(difference) = 6.97$, $t(42) = 2.43$, $p = .02$. The study by Noluck (2008)

> found no statistically significant difference between the two treatment means: *M(difference)* = 2.23, *SD(difference)* = 7.59, $t(34) = 1.25$, $p = .22$.

What would you conclude? Are the two studies giving consistent or inconsistent messages? Is the new treatment effective? What conclusions about the two studies would you expect to read in the next couple of sentences of the article?

Here are three possible answers:

1. ***Inconsistent*** "The Lucky result is clearly statistically significant at the .05 level, whereas the Noluck result is clearly not statistically significant. The two results conflict. We can't say whether the treatment is effective, and we should examine the two studies to try to find out why one found an effect and the other didn't. Further research is required to investigate why the treatment works in some cases, but not others."
2. ***Equivocal*** "One result is statistically significant, and the other statistically nonsignificant, although the two are in the same direction. We have equivocal findings and can't say whether the treatment is effective. Further research is required."
3. ***Consistent*** "The two results are in the same direction, and the size of the mean difference is fairly similar in the two studies. The two studies therefore reinforce each other, even though one is statistically significant and the other is not. The two results are consistent and, considered together, provide fairly strong evidence that the treatment is effective."

Choose which of the three answers is closest to your own opinion. You could also consider what conclusion would be most likely if such results were discussed in whatever discipline you are most familiar with. I invite you to note down your answers before reading on. Take some time; maybe have a coffee.

Second Presentation: Confidence Intervals

Now suppose that, instead, the introduction to the article described the two studies as follows:

> Only two studies have evaluated the therapeutic effectiveness of a new treatment for insomnia. Lucky (2008) used two independent groups each of size $N = 22$, and Noluck (2008) used two groups each with $N = 18$. Figure 1.1 reports for each study the difference between the means for the new treatment and the current treatment, with the 95% confidence interval on that difference.

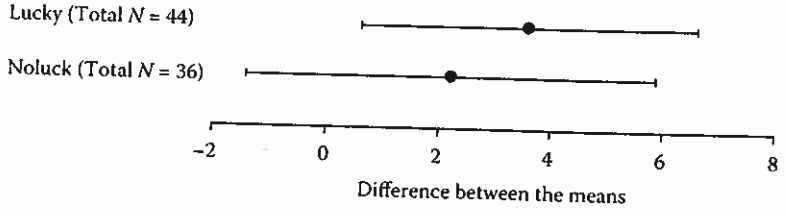


FIGURE 1.1
Difference between the means (mean for new treatment minus mean for current treatment) for treatments for insomnia in the Lucky (2008) and Noluck (2008) studies, with 95% confidence intervals. A positive difference indicates an advantage for the new treatment.

Once again, take a minute to think what you would conclude. Are the two studies giving similar or different messages? Is the new treatment effective? Choose *Inconsistent*, *Equivocal*, or *Consistent* as coming closest to your opinion. Again, note down your answers.

Third Presentation: Meta-Analysis

Now, finally, suppose the introduction to the article described the two studies, then in Figure 1.2 reported the result of a meta-analysis of the two sets of results. (Chapters 7, 8, and 9 discuss meta-analysis. Think of it as a systematic way to combine the results from two or more related studies.)

> Only two studies have evaluated the therapeutic effectiveness of a new treatment for insomnia. Lucky (2008) used two independent groups each of size $N = 22$, and Noluck (2008) used two groups each with $N = 18$. Each study reported the difference between the means for the new treatment and the current treatment.

Once again, think how you would interpret this result. Is the new treatment effective? How strong is the evidence? Again, note down your answers.

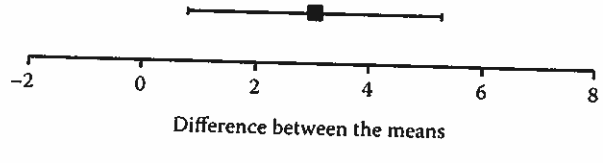


FIGURE 1.2
Difference between the means (mean for new treatment minus mean for current treatment) for treatments for insomnia, with its 95% confidence interval, from a meta-analysis of two studies that compared a new treatment with the current treatment. Total $N = 80$. A positive difference indicates an advantage for the new treatment. The null hypothesis of no difference was rejected, $p = .008$.

Some Terminology

I need to introduce some terminology, which may or may not be familiar to you. *Statistical inference* is the drawing of conclusions about the world (more specifically: about some population) from our sample data. In the Lucky–Noluck example we can assume that the *population* of interest is the notional set of all possible differences in scores between the current and new treatments, for all people affected by insomnia. A researcher should define the population carefully (Which people? What definition of insomnia?), although often the reader is left to assume. We wish to make a statistical inference about the population *parameter* of interest, which here is the mean difference in scores.

Statistical inference draws conclusions about a population, based on sample data.

You are probably familiar with *null hypothesis significance testing* (NHST), which is the most common way to carry out statistical inference in a range of disciplines, including many social and behavioral science disciplines, and some biosciences. I'll refer to disciplines that rely at least to a moderate extent on NHST as *NHST disciplines*. To use NHST you typically specify a null hypothesis then calculate a p value, which you use to decide whether to reject or not reject the null hypothesis at some significance level, most commonly .05 or .01. The first presentation of the Lucky–Noluck results used an NHST format, and the conclusion whether or not to reject the hypothesis that the population mean difference is zero hinged on whether or not p was less than .05. In other words, a result was declared statistically significant or not, depending on whether $p < .05$ or not.

Null hypothesis significance testing (N tests a null hypothesis—usually that th no difference—and uses the p value to reject or not reject that hypothesis.

A second approach to statistical inference is *estimation*, which focuses on finding the best *point estimate* of the population parameter that's of greatest interest; it also gives an *interval estimate* of that parameter, to signal how close our point estimate is likely to be to the population value. The second presentation of the Lucky–Noluck results used an estimation format. Figure 1.1 gives point estimates, which are the sample mean differences marked with the round dots. These are our best estimates, one from each study, of the true mean advantage of the new treatment. The 95% CIs are the interval estimates, and the fact that they are so long signals that our point estimates give only imprecise information about the population values, although any interpretation of CI length requires understanding of the scale we're using.

Estimation is a second approach to statistical inference. It uses sample data to calculate *point estimates* and *interval estimates* of population parameters.

The third presentation, in Figure 1.2, combines the Lucky and Noluck results. I calculated it using *meta-analysis*, which is a collection of techniques for the quantitative analysis of the results from two or more studies. At its simplest, it gives a point estimate that is a weighted average of the separate study means. (The weights depend on the sample sizes and variances of the separate studies.) It also gives an overall interval estimate that signals how precise an estimate the weighted average is likely to be.

Meta-analysis is a set of techniques for the quantitative analysis of results from two or more studies on the same or similar issues.

The Best Interpretation of Lucky–Noluck

All three presentations were based on exactly the same data. Therefore, whatever your interpretations, they should have been the same for all three presentations. Figure 1.1 indicates most clearly that the most justifiable interpretation is *Consistent*—we could even call it the correct answer. The CIs overlap very substantially, and if NHST is used to test whether there is any difference between the two studies, $p = .55$, so the two studies are as similar as could reasonably be expected even if the second was just a repetition of the first. The Lucky and Noluck results are entirely consistent, and therefore reinforce each other. Figure 1.2 shows the extent of reinforcement, where the CI resulting from the meta-analysis is shorter, indicating more precise estimation of the effect of the new treatment. Also the p value of .008 for the combined results would conventionally be taken as fairly strong evidence that the new treatment is more effective.

In later chapters we'll explore the definition and calculation of CIs and discuss several ways to think about them. For the moment it's worth noting two of those ways.

First, a CI indicates a range of values that, given the data, are *plausible* for the population parameter being estimated. Values outside the interval are relatively implausible. Any value in the interval could reasonably be the true value, and so the shorter the interval the better.

Second, CIs can, if you want, be used to carry out NHST. If zero lies within a CI, zero is a plausible value for the true effect, and so the null hypothesis is not rejected. Alternatively, if zero is outside the interval, zero is not so plausible a true value, and the null is rejected. If the intervals in Figure 1.1 are used for NHST, the results match those given in the NHST presentation: The Lucky CI does not include zero, which indicates $p < .05$ and a statistically significant result. The Noluck CI includes zero, which indicates $p > .05$ and statistical nonsignificance. CIs can easily be used to carry out NHST, although doing so ignores much of the valuable information they provide.

The New Statistics Debate

At this point I invite you to reflect on the three presentations of the Lucky–Noluck data. How did you think about each? In each case, what language would you have chosen to describe the results and conclusions, and would that language have varied over the three formats? Did NHST suggest *Inconsistent* or *Equivocal*, but the CI format *Consistent*? Did the different formats suggest that different conclusions are warranted, even though you knew the underlying data were the same, and so conclusions should all be the same? Also, were you interested in knowing how effective the new treatment is, or only whether or not it worked?

In discussions of the three presentations of Lucky–Noluck with groups of researchers, I always ask what they consider the most likely interpretations their disciplinary colleagues would make, or what they would expect to see in the journals with which they are most familiar. The most common opinion, from a number of disciplines, is that NHST is quite likely to suggest *Inconsistent*, and the CI format to suggest *Consistent*. That's only anecdotal evidence—I describe better evidence below—but it supports my contention that Lucky–Noluck illustrates how different ways of presenting results can prompt dramatically different conclusions. NHST is more likely to prompt unjustified conclusions of inconsistency or disagreement, whereas CIs may prompt more justified conclusions of consistency or similarity, at least for the simple pattern of results we've been discussing.

The Lucky–Noluck example is a small illustration of the advantages of estimation over NHST. By *the new statistics* I am referring to a shift from reliance on NHST to the use of estimation wherever possible, and also of meta-analysis whenever it can help. There are further statistical techniques that have great value and can help researchers shift from NHST, but in this book I focus on estimation and meta-analysis. I should quickly say these techniques are hardly new: CIs have been around for almost a century, and meta-analysis for several decades. It's the widespread and routine use of these techniques in disciplines traditionally reliant on NHST that would be new.

The new statistics refer to estimation, meta-analysis, and other techniques that help researchers shift emphasis from NHST. The techniques are not new and are routinely used in some disciplines, but for the NHST disciplines, their use would be new and a beneficial change.

Perhaps you are feeling unsettled by my conclusion that *Consistent* is the best interpretation, and wish to dispute it. You may be thinking there's not enough information to justify "fairly strong evidence" of an effect? Or that, although the results of the two studies are similar, it's an artificial situation, with results falling just either side of the significance boundary, and it's petty of me to criticize NHST for anomalies when results happen to fall close to the

refers to the "false clarity" of a definite decision or classification that humans clutch at, even when the situation is uncertain. To adopt the new statistics we may need to overcome a built-in preference for certainty, but our reward could be a better appreciation of the uncertainty inherent in our data.

Estimation Thinking

The most salient feature of Figure 1.1 is the CIs. The point estimates are the dots, and the intervals indicate the uncertainty of those point estimates. Figure 1.1 permits dichotomous thinking—if the intervals are used merely for NHST—but there is no reason to limit their use in this way. CIs offer much more, provided we can move beyond dichotomous thinking and adopt *estimation thinking*. Estimation thinking focuses on how big an effect is; knowing this is usually more valuable than knowing whether or not the effect is zero, which is the focus of dichotomous thinking. Estimation thinking prompts us to plan an experiment to address the question, "How much ...?" or "To what extent ...?," rather than only the dichotomous NHST question, "Is there an effect?"

Estimation thinking focuses on "how much," by focusing on point and interval estimates.

Meta-Analytic Thinking

One realization prompted by the use of CIs is that in most single studies the uncertainty is, alas, larger than we'd thought. CIs are in practice usually wider than we'd like, so we usually need to combine evidence from a number of studies. Meta-analysis gives us tools to do that, and so meta-analysis is a vital component in the new statistics. *Meta-analytic thinking* is the consideration of any result in relation to previous results on the same or similar questions, and awareness that combination with future results is likely to be valuable. Meta-analytic thinking is the application of estimation thinking to more than a single study. It prompts us to seek meta-analysis of previous related studies at the planning stage of research, then to report our results in a way that makes it easy to include them in future meta-analyses. Meta-analytic thinking is a type of estimation thinking, because it, too, focuses on estimates and uncertainty. Cumulation of evidence over studies, by meta-analysis, usually gives a more precise estimate, signaled by a shorter CI. That's excellent news, because more precise estimates are best.

Meta-analytic thinking is estimation thinking that considers any result in the context of past and potential future results on the same question. It focuses on the cumulation of evidence over studies.

The Natural Statistics?

I'm arguing that we should move, as much as practical, from NHST to the new statistics, and from dichotomous thinking to estimation and meta-analytic thinking. At one level these are drastic changes, and I don't underestimate the difficulty of breaking the hold of *p* values and dichotomous thinking. More fundamentally, however, I suspect that the natural way we think about results is (1) to focus on the most direct answer they give to our research question, (2) to consider how precise that answer is, and (3) to think how our answer relates to other results. For an example of the three steps, think of the Lucky and Noluck means and CIs, and the relation between the two results. Based on the three steps, we use our judgment to interpret the Lucky–Noluck results and draw conclusions. But that's simply a description of the new statistics in action! I suspect the new statistics may license us to think about our results in ways we'll recognize as natural, and perhaps the ways we've secretly been thinking about them all along, even as we calculate and publish *p* values. If my hunch is correct—and it needs to be examined experimentally—then once we've overcome Dawkins' tyranny and van Deemter's false clarity to move beyond dichotomous thinking, we may find that the new statistics feel rather natural. Adopting the new statistics may not feel like shifting to a different world, but as a release from restrictions and arrival at a somewhat familiar place. That place is already inhabited by disciplines that make comparatively little use of NHST, including, for example, physics and chemistry. In addition, NHST and the new statistics are based on the same underlying statistical theory, so in this way as well the new statistics may feel familiar.

Up to this point I've used the Lucky–Noluck example to describe three different ways to present results and, correspondingly, three different ways of thinking. I've also used the example to argue that estimation and meta-analysis can do a better job of statistical inference than NHST, and therefore researchers should, where possible, adopt the new statistics. Most fundamentally, the new statistics are simply more informative, but I would like, in addition, to be able to cite evidence about how researchers understand NHST and estimation. I'll shortly report some cognitive evidence, but first I'll introduce the ideas of evidence-based practice in statistics and statistical cognition.

Evidence-Based Practice in Statistics

Professional practitioners in medicine, health sciences, psychology, and many other fields strive for evidence-based practice. They should be able to

justify their treatment recommendation by referring to research showing that the treatment is likely to prove most effective in the circumstances. Clients, legislators, and the community increasingly expect nothing less, and in some cases any other approach may be ethically questionable. I suggest that choice of statistical practices should similarly be evidence based. Relevant evidence is of at least two kinds: statistical and cognitive. The technical discipline of statistics is concerned with studying statistical models and techniques, and with assessing evidence of how suitable they are for a particular situation. Such statistical evidence is undoubtedly important.

The second type of evidence, cognitive evidence, may be just as important if misconception is to be avoided, and readers are to understand results as well as possible.

Statistical Cognition

Statistical cognition is concerned with obtaining cognitive evidence about various statistical techniques and ways to present data. It's certainly important to choose an appropriate statistical model, use the correct formulas, and carry out accurate calculations. It's also important, however, to focus on understanding, and to consider statistics as communication between researchers and readers. How do researchers think about their results; how do they summarize, present, and interpret data; and how do readers understand what they read? These are cognitive questions, and statistical cognition is the research field that studies such questions (Beyth-Marom, Fidler, & Cumming, 2008). As I discuss in Chapter 2, statistical cognition has gathered evidence about severe and widespread misconceptions of NHST, and the poor decision making that accompanies NHST. It is beginning to study the new statistics, and will increasingly be able to advise how new statistical practices should be refined and used. In addition, studies of how best to teach and learn particular statistical concepts should also be helpful in guiding adoption of the new statistics.

Statistical cognition is the empirical study of how people understand, and misunderstand, statistical concepts and presentations.

I want to encourage the evidence-based practice of statistics and so, wherever possible, I'll support recommendations in this book by including boxes with brief accounts of cognitive evidence relevant to the statistical issues being discussed. Often, however, no such evidence is available. Statistical cognition is a small research field, with many vital outstanding questions. Especially in relation to the new statistics there are many interesting cognitive issues that need to be investigated. Occasionally I'll mention some. If you are interested in cognition or learning, and are looking for a research project—or even a research career—you might consider these. Please feel warmly encouraged. This research is important and could be widely influential.

How Do Researchers Think?

Box 1.1 describes my research group's investigation of how leading researchers in three different disciplines interpret results presented in NHST or CI formats, as in the first two Lucky–Noluck presentations. We examined whether NHST might prompt dichotomous thinking, and CIs might prompt estimation thinking.

As often happens in research, we were surprised. As we expected, interpretation of the CI presentation was better on average than interpretation of NHST, but the difference was small. The most striking finding was that those who saw CI results, as in Figure 1.1, tended to split, with some using NHST in their interpretation and others not. Those who used NHST said things like "one's [statistically] significant but the other isn't," and they made such comments even though NHST was not mentioned in the results they saw. Those respondents tended to see the Lucky and Noluck results as different. On the other hand, those who didn't mention NHST said things like "the intervals overlap a lot," and almost all saw the results as similar, a much better interpretation. Yes, the new statistics prompt better interpretation, but you need to think in estimation terms. If you use CIs merely to carry out NHST, you waste much of their potential and may misinterpret experimental results. Estimation thinking beats dichotomous thinking, but merely using CIs doesn't guarantee estimation thinking.

Writing Your Take-Home Messages

At the end of each chapter I suggest take-home messages that express what I think are the chapter's main points. However, it's much better for you to write your own, rather than merely read mine. Therefore, I now invite you to pause, take a coffee break, think (or look) back over the chapter, and write your take-home messages. As possible hints I'll mention that this first chapter used the Lucky–Noluck example to illustrate how different ways of presenting results can prompt different ways of thinking and different interpretations. I argued that estimation can give better interpretation than NHST. However, it's not just the choice of statistical technique that matters, but the underlying thinking adopted by the reader and the researcher. Estimation thinking, which is fundamental to the new statistics, is likely to give better understanding of results than the dichotomous thinking that naturally goes with NHST.

Finally, a word about ESCI. Use of ESCI is integrated into my discussion throughout the book, but in slightly different ways in different chapters. My aim is that you can readily work through the book without using ESCI

BOX 1.1 HOW RESEARCHERS INTERPRET NHST AND CIS

Coulson, Healey, Fidler, and Cumming (2010, tinyurl.com/cisbetter) investigated how researchers think about the Lucky–Noluck results, presented in NHST or CI formats, with the aim of testing the new statistics predictions that CIs elicit better interpretation. We emailed authors of articles published in leading psychology, behavioral neuroscience, or medical journals and presented results in either an NHST or a CI format. We asked researchers, "What do you feel is the main conclusion suggested by these studies?" We then asked them to rate their extent of agreement with statements that the two results are "similar" on a scale from 1 (*strongly disagree*), to 7 (*strongly agree*).

There was little sign of any differences between disciplines, so we combined the three. Overall, there was enormous variability in respondents' interpretations and ratings, with all responses from 1 to 7 being common. Recall that the two studies are actually consistent, and so the best answers are ratings of 6 or 7. The means were 3.75 for NHST and 4.41 for CI. The CI mean was greater by 0.66, 95% CI [0.11, 1.21]. (Square brackets are the CI reporting style recommended in the APA *Publication Manual* and are what I will use throughout this book.) Yes, CIs gave higher ratings, which is good news for the new statistics, even if the difference of 0.66 points on the 1–7 scale is not large. (You noticed that the difference was statistically significant, because the CI did not include zero?) However, there was enormous variability, and mean ratings were close to the middle of the scale, not around 6 or 7. Most respondents performed poorly, even though the pattern of the Lucky–Noluck results was simple and no doubt familiar to many.

We analyzed the open-ended responses to the initial question about the "main conclusion suggested by these studies." We were struck that many respondents who saw the CIs still mentioned NHST. We divided respondents into those who mentioned NHST when interpreting the CIs shown in Figure 1.1 and those who didn't. Figure 1.3 shows for those two groups of respondents the percentages of responses that described the Lucky and Noluck results as similar or consistent, or as different or inconsistent. Those who mentioned NHST were likely (33/55, or .60) to consider the results "different," as dichotomous thinking would suggest. In striking contrast, of those who avoided any reference to NHST almost all (54/57, or .95) gave a better answer by rating the results "similar," as estimation thinking would suggest. In other words, most who mentioned NHST gave an incorrect answer, whereas almost all who did not mention NHST gave a correct answer. The two proportions of respondents who answered correctly by saying "similar" were 22/55, or .40, and

54/57, or .95, for the two groups of respondents, respectively. The difference between those two proportions is .55, [.39, .67]. (In Chapter 14 we'll use ESCI to find such a CI on the difference between two proportions.) In a second study we found evidence supporting this result. Our conclusion was that CIs can indeed give better interpretation, but only if you adopt estimation thinking and regard them as intervals, and avoid merely using them to carry out NHST.

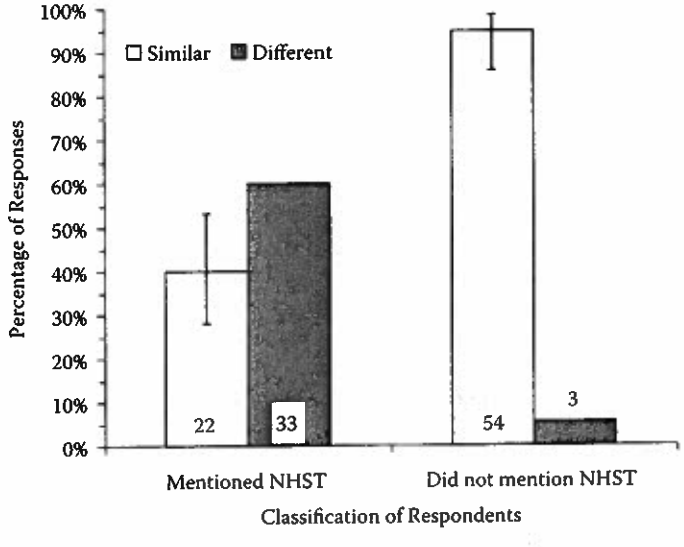


FIGURE 1.3
Percentage of open-ended responses classified as indicating the Lucky and Noluck studies gave "Similar" or "Different" results, for respondents who "Mentioned NHST" or "Did not mention NHST." Error bars are 95% CIs. Numbers of respondents are shown at the bottom. (Adapted with permission from M. Coulson, M. Healey, F. Fidler, & G. Cumming (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Quantitative Psychology and Measurement*, 1:26, 1–9.)

There's a second part to the story. We included medical researchers because medicine has routinely reported CIs since the 1980s, although data interpretation in medical journals is still often based on NHST, even when CIs are reported. Despite their many years of experience with CIs, medical researchers did not perform better than researchers in the other disciplines. It seems that merely using CIs does not guarantee estimation thinking. This conclusion from medicine reinforces our finding that CIs can indeed give better interpretation, but you need to avoid using them just to carry out NHST.

at all, but that using the software adds interest and encourages deeper understanding. Some of the book's main messages are illustrated using simulations in ESCI that I hope make the ideas clear and memorable. In this first chapter I haven't used ESCI yet, but the Exercises will now introduce ESCI, then use it to explore the ideas in the chapter.

Hints for ESCI Exercises

In most chapters the main discussion in the text refers to ESCI, and the exercises at the end of each chapter also use it. At the back of the book there's commentary on most of the exercises. I invite you to take the following approach to using ESCI:

- Focus on understanding. Can you explain the concept to someone else, perhaps using ESCI? Can you draw a picture, make your own example, and recognize the concept when you encounter it elsewhere?
- Look out for images in ESCI that represent a concept. In many cases I hope they are useful for remembering and understanding the concept, and can serve as a signature or logo for it. These may be a picture or a movie—a running simulation. Examples to come include the *mean heap, dance of the CIs,* and diagrams showing *rules of eye.*
- Focus not on the software or the interface, but on the things that really matter—the statistical ideas.
- ESCI is intended first as a playground for understanding statistics, and second as a set of tools for calculating and presenting CIs in simple situations. Watch out for tools you might find useful in the future—perhaps for calculating CIs or for making figures with CIs to present your own data.
- ESCI exercises are at first quite detailed, but I encourage you to explore as widely as you wish, perhaps using your own data and examples. In later chapters they will be much less step-by-step, and I'll mainly give broad suggestions and invite you to use ESCI in whatever ways are most useful for you.

Exercises

1.1 Load and run **ESCI chapters 1–4**. Appendix A can assist.

1.2 Click the bottom tab to go to the page **Two studies**. Compare with Figure 1.4.

decision boundary? In fact, however, CIs provide more and better information than NHST in almost all situations, and not only for Lucky–Noluck.

You might also argue that NHST gives a basis for clear decisions, and often the world needs decisions. After all, the practitioner must decide either to use a therapy with a client, or not; and the regulatory authority must either approve a drug or refuse approval. Yes, we do need to make decisions, but NHST often gives misleading guidance about decisions, as it did for Lucky–Noluck, whereas estimation and meta-analysis present all the available evidence and are thus most informative for making decisions.

Alternatively, you could argue that I'm criticizing not so much NHST as the way some people use it. Appropriate use of NHST with Lucky–Noluck would perhaps involve using NHST to compare the two—and find no statistically significant difference between them. Or perhaps NHST would be applied only to the result of the meta-analysis combining the two. Those are reasonable points to make, but my most fundamental criticism of NHST is not about the way it's used, but that it gives such an incomplete picture, in contrast to the full information provided by estimation. In addition, I'll report evidence that many researchers don't understand NHST correctly and don't use it appropriately. However, even if those problems could be overcome, the more fundamental shortcomings of NHST remain.

You could also protest that we don't know how to interpret the scale of measurement: A difference of 3 may be trivial, and of little practical or scientific value. That's true, and is an important issue we'll discuss in Chapter 2, but it doesn't undermine the conclusion that we have fairly strong evidence of some difference and, most usefully, a numerical estimate of its size—whether we judge that large or small, important or trivial. This book is mainly about the new statistics and how to use them in practice, but the comparison with NHST will come up in several places. In Chapter 15 I'll discuss in more detail a number of queries a skeptic could raise about the new statistics, and give my answers.

In the next chapter I'll discuss further the problems of NHST and describe how statistical reformers have, for more than half a century, been publishing critiques of NHST and advocating a shift to estimation or other techniques. The prospects for achieving real change may now be better than ever before, because the sixth edition of the *Publication Manual of the American Psychological Association* (APA, 2010) strongly recommends CIs, specifies a format for reporting them, and gives many examples. Like earlier editions, it also gives NHST examples, but its detailed guidance for estimation is new. The *Manual* is used by more than 1,000 journals across numerous disciplines—way beyond just psychology—and every year enormous numbers of students learn its rules of style. The new edition states unambiguously: "Wherever possible, base discussion and interpretation of results on point and interval estimates" (p. 34). This is a strong

endorsement of the new statistics, which I hope will be influential and lea to improvements in the way research in many disciplines is conducted.

Now I want to discuss three different ways of thinking that are relate to my three presentation formats described previously, and which seen to me fundamental to all consideration of statistical reform and the ne statistics. Changing our habits of thought may be one of the biggest cha lenges of moving to the new statistics, but also potentially a valuable out come of making the change.

Three Ways of Thinking

Dichotomous Thinking

The first presentation format of Lucky–Noluck emphasizes dichotomou decision making: NHST results in a decision that the null hypothesi is rejected, or not rejected. Such dichotomous decision making seem likely to prompt *dichotomous thinking* which is a tendency to see the world in a either–or way. Experiments are planned hypotheses formulated, and results ana lyzed, all within a framework of two com pletely opposed possibilities: A result i statistically significant or not. The first Lucky–Noluck presentation report for each study the mean (M) and standard deviation (SD) of the differ ences, and values of t and p. Our point estimate of the difference betwee the new treatment and the old is M. Yes, M is reported, but NHST habit may prompt us to skim through the text searching mainly for the p value Small p usually means success and may elicit joy; large p may elicit disap pointment and frustration. Such an either–or outcome goes with dichoto mous thinking: We formulate our research in terms of a null hypothesis o zero effect, and finish with a dichotomous decision that we can, or cannot reject it. *Dichotomy* comes from the Greek "to cut in two," and NHST give us just two distinct options for decision. It seems plausible that dichoto mous thinking and use of NHST are mutually reinforcing. If so, dichoto mous thinking may be an obstacle to adoption of the new statistics.

Dichotomous thinking focuses on making a choice between two mutually exclusive alternatives. The dichotomous reject-or-don't-reject decisions of NHST tend to elicit dichotomous thinking.

Why does dichotomous thinking persist? One reason may be an inher ent preference for certainty. Evolutionary biologist Richard Dawkin (2004) argues that humans often seek the reassurance of an either–o classification. He calls this "the tyranny of the discontinuous mind" (p. 252). Computer scientist and philosopher Kees van Deemter (2010, p. 6

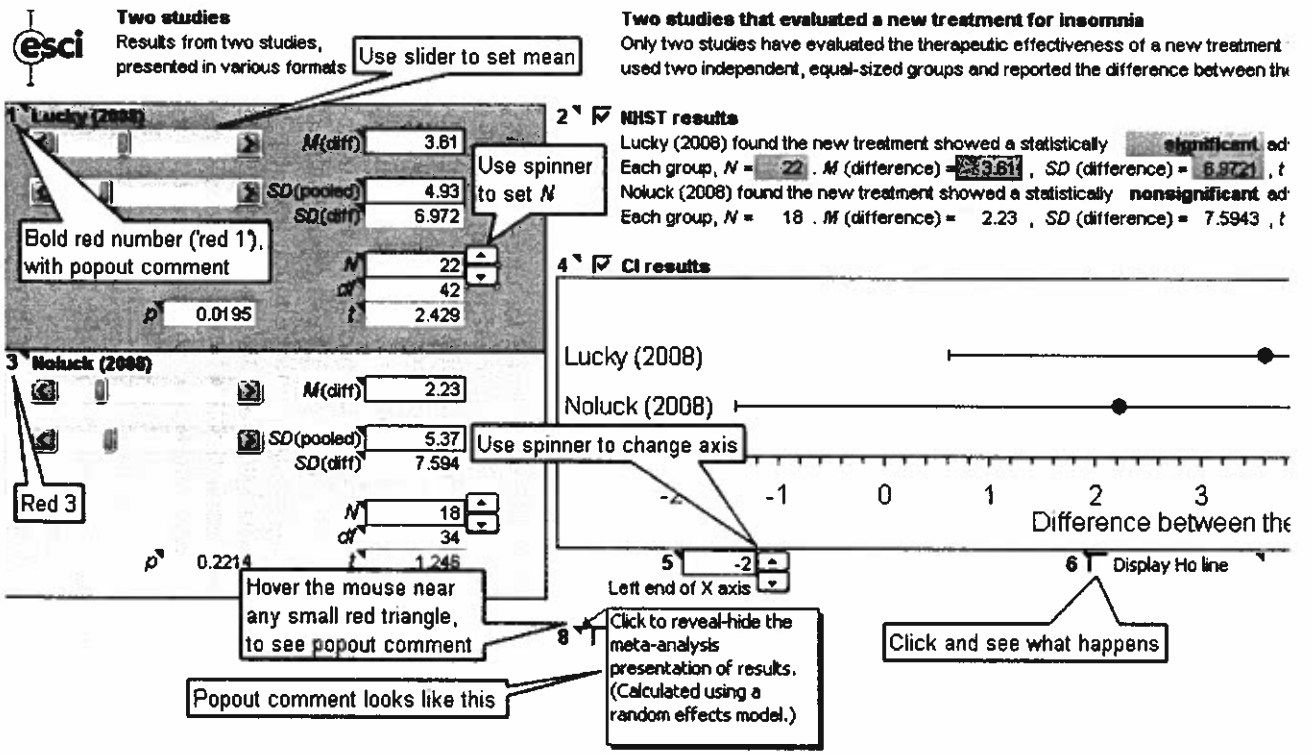


FIGURE 1.4
A partial screen image from the **Two studies** page of **ESCI chapters 1–4**. The Lucky (2008) and Noluck (2008) values have been set to match the first presentation, at the start of this chapter. The callouts point to some features of the display.

1.3 Find red 1 (the bold red 1 near the top left) and read the popout (hover the mouse near the little red triangle). A "slider" looks like a horizontal scroll-bar, and a "spinner" has small up and down arrowheads you can click.

1.4 Use the top slider to set *M*(diff) = 3.61 for the Lucky study. Use the next slider to set *SD*(diff) as close as possible to 6.97. (The slider actually sets the pooled SD within groups, and *SD*(diff) is calculated from that.) Use the spinner to set *N* = 22 for Lucky.

If you ever seem to be in an ESCI mess, look for some helpful popout comments. Appendix A may help. Or experiment with the controls—you won't break anything—and see whether you can straighten things out. Or you can close the Excel module (don't Save), then reopen it to start again.

1.5 Note the description of the Lucky results at red 2. This is the NHST version. Check that it matches what you expect.

1.6 At red 3, set the values from the first presentation for Noluck, including *N* = 18. Are the results as you expect?

1.7 Click at red 4 to see the CI results. Do they match Figure 1.1? Reveal the meta-analysis results. Are these also as you expect?

Excel macros are needed to make that work. If you cannot see the figure with CIs, you may not have enabled macros. See Appendix A.

1.8 Hide the CI results (click at red 4) and meta-analysis results. Focus on *p* for Lucky. How do you think this *p* would change if

- You increase *M*(diff); you decrease it?
- You increase *SD*(diff); you decrease it?
- You increase *N*; you decrease it?

1.9 Play with the controls at red 1 to see whether your predictions about the *p* value for Lucky were correct.

There is evidence that making your own predictions like this, then experimenting to test them out, can be an effective way to cement understanding. Did you actually make the predictions before you played? It's worth making your best predictions and writing them down, before you start experimenting. That's more effective use of your time.

1.10 Reveal results presented in the CI and meta-analysis formats (click at red 4 and red 8). Play around with the sliders and spinners for Lucky and Noluck (at red 1 and red 3), and watch what happens in all three formats. Look for relationships between the formats. Make predictions then test them.

Lots of this book's exercises are like the last one: fairly open ended. They are probably of most use if you adopt strategies such as the following: Take time to explore, collaborate with someone else, set challenges for yourself or others, and write down your conclusions and questions. Where possible, try to find parallels with other statistics books with which you are familiar.

1.11 Focus on the Lucky CI in the CI figure (red 4). Click at red 6 to mark the null hypothesis. Adjust *M*, *SD*, and *N*, and note how *p* changes and whether the CI covers zero. What is *p* when the CI includes zero? When it misses zero? When the lower limit (LL) of the CI just touches zero?

1.12 Reveal the meta-analysis results. Play with the spinners at red 5 and red 7—to the right of the display, not shown in Figure 1.4—to see how you can control the horizontal axis in the CI and meta-analysis figures.

1.13 By now you have used every control on this page. Browse around the page and read the popouts, wherever you see little red triangles. Is everything as you expect? Does it make sense?

1.14 Play with the values for Lucky and Noluck, and watch how the meta-analysis result changes. What is the relation between the *p* value for the meta-analysis result, which is shown just below the meta-analysis figure, and whether the CI crosses zero?

1.15 Watch how the meta-analysis CI relates to the separate CIs for the two studies. Think of the meta-analysis as combining the evidence from the two studies, as expressed by the separate CIs. Usually, the meta-analysis CI will be shorter than each of the separate CIs. Does that make sense?

You can talk about CIs as being long or short, or equivalently as being wide or narrow. Either pair of terms is fine.

1.16 Make the means for Lucky and Noluck very different, so the two CIs don't overlap. Would you regard the two studies as giving *Consistent* or *Inconsistent* results? How long is the meta-analysis CI? Does that make sense?

1.17 Play around further, then write down some conclusions of your own. Do you prefer long or short CIs? Are you comfortable examining a CI figure and deciding whether the result is statistically significant or not?

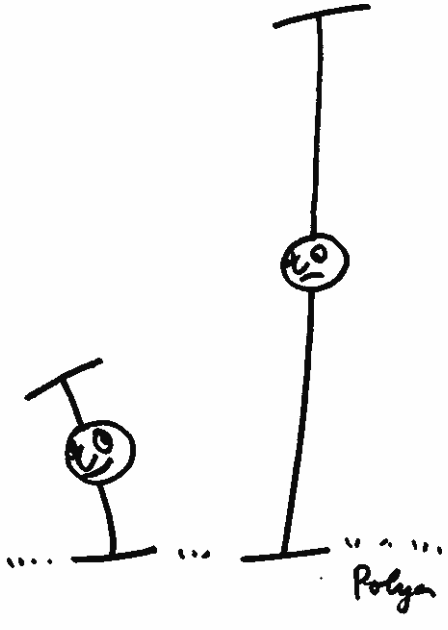
1.18 Think how you could use this page to invent games to challenge your friends—perhaps show them one format and have them predict what another will look like.

1.19 As you inspect the three formats, do they prompt the three different ways of thinking? To which of the presentations do you find yourself returning? Which seems to give better insight? Can you recognize which type of thinking you are using at any particular moment?

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

That takes time and effort, but the evidence is that generating your own summary is worthwhile and more valuable than merely reading mine.

Sometimes my "take-home messages" include a "take-home picture" or a "take-home movie." These are images, static or moving, that I hope provide vivid mnemonics for a concept and help the concept make intuitive sense. Often, but not always, they come from ESCI. They've become sufficiently embedded in your thinking when you dream about them.



Take-Home Messages

- The way results are presented really matters. Change the format and the researcher may interpret them differently, and a reader may receive a different message.
- Assessing what message a presentation format conveys is a cognitive question that the research field of statistical cognition seeks to answer.
- Evidence-based practice in statistics is desirable, and cognitive evidence can help.
- CIs are more likely to give a better interpretation of results than an NHST format, at least for the frequently occurring pattern of results shown in Figure 1.1.
- *Take-home picture*: The Lucky–Noluck pattern of Figure 1.1. This figure illustrates that a large overlap of CIs can indicate consistency of results, even when one CI includes zero and the other doesn't, so that NHST suggests, misleadingly, that the two studies give inconsistent results.
- NHST may prompt dichotomous thinking, whereas CIs are likely to encourage estimation thinking and meta-analytic thinking.
- The fundamental advantage of estimation, CIs, and meta-analysis is that they provide much fuller information than NHST, which focuses on the very limited question, "Is there an effect?"
- Merely using CIs may not suffice to overcome dichotomous thinking. In addition, CIs should be interpreted as intervals, with no reference to NHST.
- *The new statistics* aim to switch emphasis from NHST to CIs and meta-analysis, and from dichotomous thinking to estimation thinking and meta-analytic thinking.
- NHST disciplines should be able to improve their research by progressively moving to the new statistics.