4

Confidence Intervals, Error Bars, and p Values

This chapter presents more about CIs, including some beautiful pictures and two further approaches to CI interpretation. The main issues are

- Error bars, and how they can show various types of information in figures
- Cat's-eye pictures that show the beautiful shape of a CI
- The fourth approach to interpreting CIs—in terms of their shape
- The relation between CIs and *p* values
- The fifth approach to interpreting CIs—with reference to *p* values
- One-sided CIs, which correspond to one-tailed NHST

The Error Bar, a Picture With a Dozen Meanings

Does Figure 4.1 show the mean number of ice creams consumed by 10-year-olds or the median response time to a red stoplight? What do the error bars represent? These are good questions, and Figure 4.1 fails to provide answers. I'll refer to the simple graphic shown on the column and dot in Figure 4.1 as *error bars*, or sometimes simply *bars*. Error bars define a range of values around a point estimate such as a mean. The trouble is that bars can be used to depict various different types of ranges.

Error bars, or *bars*, are a simple graphic that marks an interval around a mean or other point in a figure.

When you see Figure 4.1, what questions spring to mind? Probably the most basic concern is the dependent variable, what it measures, and what its values mean. Ice creams or braking times? Then we need to know what the column and the big dot are reporting—perhaps the sample mean, a median, or a frequency? Labels on the figure or the figure caption needs to give clear answers to these questions.

Seeing the error bars should prompt an additional question, because, although the error bar graphic is familiar, it is, unfortunately, ambiguous.

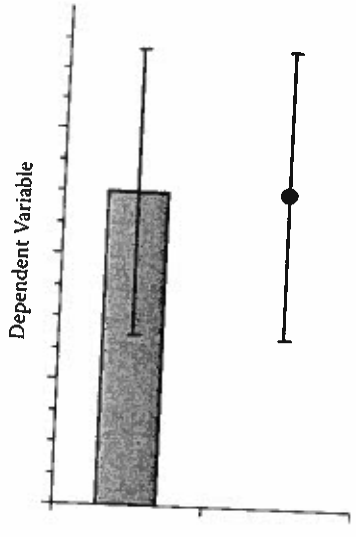


FIGURE 4.1
Two ways to display a mean, or other descriptive statistic, with error bars.

Do the bars show a CI? Or the SE? Or something else? We could refe only slightly histrionically, to the *tragedy of the error bar,* which is that bar don't automatically state what they are reporting. Seeing error bars need to prompt questions about what the represent. Alas, figures often fail to stat what the bars represent, in which case it impossible to make sense of the figure.

The *tragedy of the error bar* is how I refer to the unfortunate fact that error bars don't automatically announce what they represent. We need to be told.

Column or Dot?

Error bars represent some measure of variability or uncertainty, but even the way the point estimate is depicted may influence our perception of variability. Referring to Figure 4.1, consider a further question: Column or dot? Both are common, but which prompts better interpretation? If the vertical axis starts at zero, the height of a column gives a direct representation of effect size that can be useful for appreciating the effect. On the other hand, a column has a sharply defined top end. Yes, we can calculate the sample mean as precisely as we like and picture it with a sharp-topped column. But one main message of this book is that sampling variability is often greater than we suspect. A sharp-topped column may hint that we have precise information about the population. Even if the suggestion is subliminal, that would be misleading. On the other hand, a dot, as in Figure 4.1, reports the sample mean clearly, but may possibly not

give such a strong hint about precision. Research is needed to investigate these speculations. In the meantime, I'll generally prefer dots to columns. In either case, however, error bars should be used to provide explicit information about variability or uncertainty.

While we're thinking about columns and dots, I'll mention *Graph Design for the Eye and Mind*, by Stephen Kosslyn (2006). It's a book based on statistical cognition and other research about how people interpret—or misinterpret—graphs. Kosslyn uses the research findings to formulate good advice about how to design figures to report data. He discusses research on the column and dot issue (see pp. 46–53), but doesn't consider what, for me, is the vital question—to what extent does each give an appreciation of the uncertainty in the data? That question awaits investigation. There is much that's useful in Kosslyn's book, although there is nothing on confidence intervals and only a few mentions of error bars.

Confidence Intervals and SE Bars

The most common uncertainty is whether error bars represent a CI, or are SE bars, where *SE bars* are error bars that extend from one SE below to one SE above the mean, or other point estimate. Unfortunately, different research fields have different customs. In medicine, for example, CIs are routinely reported, and so unidentified bars are probably CIs, although they might not be. In some biological disciplines, however, SE bars are routinely shown in figures, and some researchers regard any mention of "error bars" as automatically implying SE bars. I recommend the safer policy of using the term "error bars" to refer simply to the graphic illustrated in Figure 4.1, without any assumption of a particular meaning.

Standard error bars, or *SE bars*, extend from one SE below to one SE above the mean. Unfortunately, some researchers and disciplines assume that "error bars" means SE bars.

Should we prefer CIs or SE bars? You won't be surprised to hear that I recommend CIs. To explain why, I need to discuss the role of N, the sample size, then introduce the idea of *inferential information*. Figure 4.2 illustrates 95% CIs and SE bars for three samples, of sizes 5, 20, and 80. These are random samples from the same normal population, but I tweaked them a little so they all have the same $M = 50$ and same $s = 17$. That should help comparison of the error bars, which is the aim here. For each sample, the 95% CI is on the left, and the SE bars are on the right. Overall, the two types of bars are very different, with CIs being around twice the length of SE bars, or longer. In addition, both CIs and SE bars show large changes in length with N. We'll see that the relation between length and N is different in the two cases, and that this difference underlies the advantage of CIs.

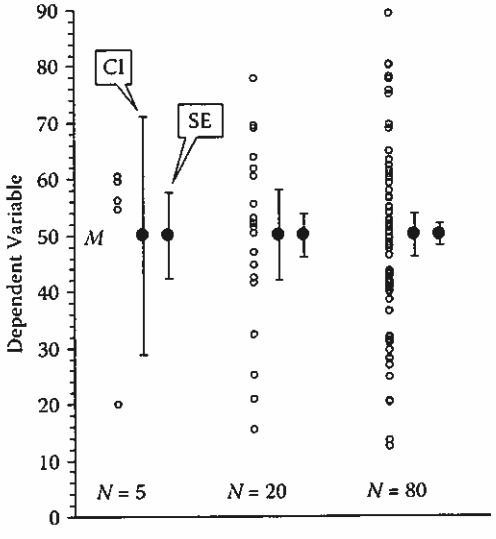


FIGURE 4.2
Three samples, of sizes 5, 20, and 80, displayed as dot plots. All samples have the sam mean $M = 50$, and same standard deviation $s = 17$. For each sample, the error bars on the le represent the 95% CI, and on the right are SE bars, which mark ± one SE.

First I'll focus on the CIs, which show an especially marked change i length across the three values of N. Using Equation (3.3) and $N = 5$, MO for the CI is

$$t_{.95}(N-1)\times s/\sqrt{N} = t_{.95}(4)\times 17/\sqrt{5} = 2.776\times 7.603 = 21.11.$$

For $N = 5$ the critical value of t is 2.776, and you can use Appendix B an ESCI **Normal z t** to check that the critical value is correct. For $N = 20$, th critical value of t is 2.093 and MOE is 7.96, and for the $N = 80$ sampl the critical value of t is 1.990 and MOE is 3.78.

Now consider the SE bars shown on the right for each sample. Becaus $SE = s/\sqrt{N}$, these bars also get shorter as N increases. There's a factor 4 increase in sample size from one sample to the next from left to rigl across Figure 4.2, which means there should be a factor of $\sqrt{4} = 2$ decreas in SE, from sample to sample. The SE bars in the figure do show th decrease: For the $N = 5$ sample, $SE = 17/\sqrt{5} = 7.60$. For the $N = 20$ sampl $SE = 3.80$ and the bars are half as long. For $N = 80$, $SE = 1.90$ and the ba are again halved in length. Do these values look about right, as you rea the SE bars in Figure 4.2?

Compare that pattern for SE bars with changes in the CIs for differe values of N. The lengths of the three CIs reflect the three MOEs, whic are 21.11, 7.96, and 3.78, from left to right—from the smallest to large

samples. Compare 21.11 and 7.96 to see that the CIs shorten by distinctly more than a factor of 2 from the $N = 5$ to the $N = 20$ samples. From the $N = 20$ to the $N = 80$ samples, the CIs shorten by close to a factor of 2.

Another approach to considering CIs and SE bars is to compare the two sets of bars for each sample. For $N = 5$, the critical value of t is 2.776, considerably larger than 2, and the CI appears considerably longer than double the length of the SE bars. However, for both the $N = 20$ and $N = 80$ samples, the 95% CI appears about double the length of the SE bars, because the critical values of t for those samples are 2.093 and 1.990, respectively—both quite close to 2. Now I need to introduce inferential information and a *rule of eye*.

Inferential Information

CIs give us *inferential information*, which is information that supports an inference about the population. It's calculated from the sample, but informs us about the underlying population. The sample mean, M, gives inferential information when we use it as our point estimate of the population mean. A CI provides inferential information because it tells us how precise our point estimate is for μ, the parameter we're estimating.

Inferential information is based on the sample data, but tells us about the population.

SE bars usually don't provide accurate inferential information. Often you can interpret SE bars as being, approximately, the 68% CI; there's more about that later in this chapter. Also, as we saw previously, you can double the length of SE bars to get, approximately, the 95% CI. That's a useful *rule of eye*, by which I mean a generally useful guideline to remember when interpreting figures (Cumming & Finch, 2005, tinyurl.com/inferencebyeye). Like a rule of thumb, it's not always exact, but it's often helpful. If N is at least 10, the rule is reasonably accurate. For $N = 10$ or less, however, it becomes progressively more in error as N decreases. Figure 4.2 illustrates that, for $N = 5$, the 95% CI is almost three times the length of the SE bars. It would be seriously inaccurate to interpret SE bars when $N = 5$ as a 68% CI, or twice their length as a 95% CI. That's the trouble with SE bars: They don't provide what we want, which is accurate inferential information. CIs by definition provide that information.

A *rule of eye*: Double the length of SE bars to get, approximately, the 95% CI. This rule is reasonably accurate for means when N is at least 10.

Why then are SE bars used so commonly, especially in some disciplines? That's an excellent question. There has been almost no study of how researchers think about or interpret SE bars, but they are probably seen as providing a type of inferential information, perhaps a rough indication of precision. Yes, you can often interpret SE bars inferentially by doubling their length to get, approximately, the 95% CI, but, as we have

seen, that strategy fails for small samples. Also, for some measures, including correlations and proportions (see Chapter 14), CIs are not calculated from the SE, so in those cases also SE bars could be misleading. The conclusion must be that CIs should be preferred to SE bars. Medicine agrees and expects researchers to report CIs. In Chapter 6 I'll mention evidence that many researchers don't appreciate the distinction between SE bars and CIs, even though they differ by a factor of about 2! Anyone who doesn't appreciate the difference might prefer SE bars because they are shorter and thus suggest less uncertainty in the data. However that's an illusion, because it's the CI that gives accurate information about uncertainty—because that's what they are designed to do. We should prefer CIs to SE bars. Simple as that.

Descriptive Information

Additional error bar confusion arises when bars are used to convey not inferential, but *descriptive information*. I need to leave error bars for a moment, and say something about descriptive information, which, as you would guess, describes the sample data. It may provide a complete description, like the dot plots in Figure 4.2, which mark every data point, or it may be a descriptive statistic that summarizes an important aspect of the sample. In Figure 4.1 the column and dot very likely mark the sample mean, which is the most common descriptive statistic as well as being a point estimate that provides inferential information. Other descriptive statistics are *s*, the median, and the range. You may know about boxplots and frequency histograms, which are descriptive pictures of a sample. All of these give us information about the data points in the sample, their values, and how they are spread.

Descriptive information tells us about sample data.

I've been referring to the set of data as "the sample," and almost always in this book we'll discuss data sets that are random samples from a population, which is our real interest. However, there are other data sets. You might be investigating the world's top 100 performers in your favorite sport. You could plot data showing their performance times or their earnings. Your investigation would be based on a variety of descriptive statistics and pictures, which tell you about those 100 sports people. There's no thought of the data being a random sample from some larger population—it's descriptive information about those 100 people that fascinates you.

There's another reason to be interested in descriptive information. To introduce it, I'll mention Wilkinson and the Taskforce on Statistical Inference (1999, tinyurl.com/tfsi1999), which is the report from a group of statistical experts set up by the APA. I strongly recommend this report, which included much wonderful and down-to-earth advice, including this statement: "As soon as you have collected your data, before you comput

any statistics, *look at your data*" (p. 597, emphasis in the original). Examining descriptive displays and summary statistics allows you to appreciate the whole data set and identify problems, or intriguing aspects, before you launch into inferential analysis. You may be able to check assumptions. I'll say little more about this vital first step of data analysis, and will usually assume the data we're discussing have undergone this examination.

Error Bars for Descriptive Information

Error bars are sometimes used to report descriptive rather than inferential information, and it's part of the tragedy of the error bar that the same error bar graphic is used for both. Descriptive bars tell us about the spread of data points within the sample. They may indicate the range, from the lowest to highest data points, or interquartile range, but most often indicate the sample standard deviation s. Figure 4.3 is the same as Figure 4.2, but with *SD bars* also shown. The SD bars extend a distance s below M and s above M, and are the same for all the samples, being ±17 in each case. The data points might appear more widely spread in the $N = 80$ sample, and the range does increase with N, but that's because larger N makes it more likely that at least a few extreme data points will be sampled. For a given population,

SD bars describe the spread of data points in a sample. For a given population, they don't change systematically as N changes.

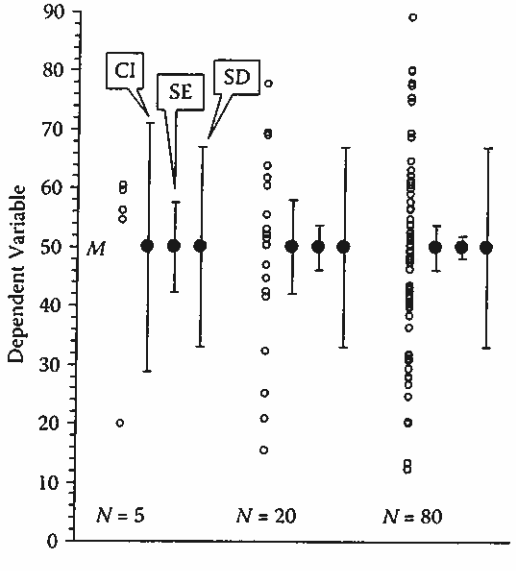


FIGURE 4.3
The same three samples as in Figure 4.2, with an additional set of error bars, on the right for each sample, that show the SD, which is $s = 17$ in every case.

we don't expect any systematic change in s if we take larger or smaller samples. Figure 4.3 illustrates how the CIs and SE bars change dramatically with N, but the SD bars don't.

A small complication is that the sample SD, like the sample mean, provides both descriptive and inferential information and, as Box 4.1 explains, s can be calculated slightly differently for those two purposes. However, I follow common practice by using the same calculation of s, with $(N - 1)$ in the denominator, whatever the purpose. Descriptively, s measures the spread of data points in the sample. Inferentially, whatever the value of N, s is our best estimate of σ, so we expect s to be roughly similar to σ, whatever the sample size. That's why the pattern of SD bars in Figure 4.3 is so different from that for the other bars. Yes, the value of s is likely to bounce around for successive samples, as we saw in Chapter 3, Exercises 3.9 and 3.10, and more so for small N. But s is, on average, close to σ, and that's true for any N. If that's all a bit confusing, focus on Figure 4.3,

BOX 4.1 TWO DIFFERENT SDS

Here's an optional extra point about SDs. As you may know, there are two different ways to calculate the SD of a data set. The first uses N in the denominator to give a descriptive statistic, the SD of the data set itself:

$$s_{\text{descriptive}} = \sqrt{\frac{\left(\sum X_i - M\right)^2}{N}} \tag{4.1}$$

You might choose that SD to describe the heights of your 100 top athletes, because you are considering that data set in its own right, and not as a sample from a population. The other formula uses $(N - 1)$ to give an inferential statistic that's the best estimate of σ:

$$s_{\text{inferential}} = \sqrt{\frac{\left(\sum X_i - M\right)^2}{N - 1}} \tag{4.2}$$

Only when N is very small is there much difference between the two. I'm going to simplify things by always using the formula with $(N - 1)$ and the symbol s, as I've been using so far. Many textbooks, software packages, and ESCI do likewise. That s is best for estimating σ, which is usually our main concern, but it's also a pretty good descriptive measure of the variation within a data set.

which summarizes the main story: CIs and SE bars vary markedly with N, whereas SD bars don't.

To show variability within a sample, SD bars can be useful, but consider using a dot plot—as in Figures 4.2 and 4.3—or a boxplot instead. The main error bar question remains CIs versus SE bars, and on this issue the bottom line is: Report CIs, and if you see SE bars, double them in your mind's eye to get approximate 95% CIs—unless the sample size is less than around 10. The further bottom line is that it's absolutely essential that every figure with error bars states clearly what the bars represent. You'd think that wouldn't need saying, but in a survey of psychology journals (Cumming et al., 2007) we found that 32% of articles that included figures with error bars did not state what the error bars represented. That's terrible because, without that information, the figure is not interpretable. To repeat: It's essential to follow the requirement of the *Manual* (APA, 2010, Chapter 5) that every figure showing error bars must state clearly what the bars represent.

Prefer CIs over SE bars. In any case, a figure showing error bars must state what the bars represent.

Many disciplines have confusions about error bars. Our article in the *Journal of Cell Biology* (Cumming, Fidler, & Vaux, 2007, tinyurl.com/errorbars101) explained SD, SE, and CIs, and offered rules of eye. It's very basic, but was a hit and was downloaded thousands of times. If you are comfortable with the different types of bars illustrated in Figure 4.3 you are ahead of many published researchers out there. Take a pat on the back.

The Shape of a Confidence Interval

I now want to look *inside* a CI and develop a picture to give us a fourth way to interpret CIs. It's a novel picture, but I hope it's helpful for understanding CIs in practical situations. Here's a preview: In Chapter 3 our second interpretation of a CI stated that values inside the CI are plausible as true values for μ, and values outside the interval are relatively implausible, but not impossible. Our fourth interpretation refines that by describing how the plausibility that a value is μ is greatest for values near M, in the center of the CI. Plausibility then drops smoothly to either end of the CI, then continues to drop further outside the CI.

The Cat's-Eye Picture

Consider Figure 4.4, which shows the dance of the CIs for samples of size $N = 15$ from a population of Hot Earth Awareness Test (HEAT) scores with

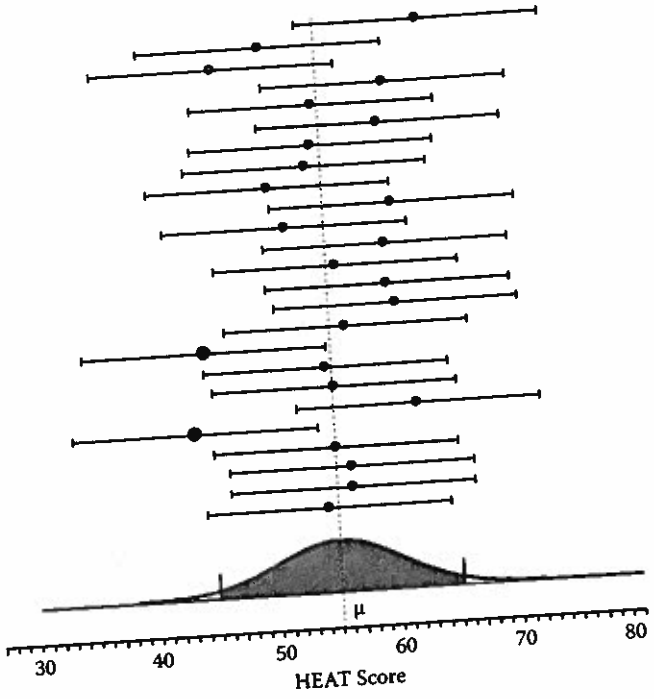


FIGURE 4.4
Dance of the 95% CIs for samples of size $N = 15$ from a normally distributed population of HEAT scores with $\mu = 55$, and $\sigma = 20$ assumed known. If a CI does not capture μ, the sample mean is shown as a large black dot. (In ESCI, those CIs would be red.) The curve at the bottom is the sampling distribution of M. It's a normal distribution and has a mean of 55 and standard deviation of $SE = 20/\sqrt{15} = 5.16$. The shaded area includes 95% of sample means, and therefore extends MOE below and above μ.

mean $\mu = 55$, which is marked by the vertical dotted line. I'll divide my discussion into four steps, but before I start, note that I speak of "where M (or its CI) falls in relation to the unknown μ." In other words, I take a sample and find that its M (or its CI) falls near to, or a little distance from, μ. That wording is a bit awkward, but emphasizes that it's M and the CI that vary, whereas μ is fixed as well as unknown. That's why, as I explained in Chapter 3, I talk about plausibility, not probability. Keep Figure 4.4 in mind, and the dance of the CIs.

1. The curve at the bottom in Figure 4.4 is the sampling distribution of the sample mean, M. The shaded area includes 95% of sample means, so we know from Chapter 3 that it extends MOE on either side of μ. The height of the curve at any HEAT value on the horizontal axis at the bottom indicates the relative likelihood that our M falls at that value. The curve is highest at $\mu = 55$, and so M is

most likely to fall at, or very near, μ. At values a little way from μ, the curve is a little lower and, correspondingly, *M* is a little less likely to fall at those values. The height of the curve decreases smoothly, and at a distance of MOE to the left or right of μ (i.e., at either end of the shaded area) its height is about one-seventh as great as it is at μ. Therefore, *M* is about seven times as likely to land at μ as it is to land at MOE below μ, or at MOE above μ. In brief, the curve tells us how values farther from μ become progressively less likely for our *M*.

2. Now consider estimation error ($M - \mu$), which we encountered in Chapter 3. It's the distance between the *M* of a particular sample and μ. I stated previously that the curve at the bottom in Figure 4.4 is the sampling distribution of *M*, but it's also the sampling distribution of estimation errors, meaning ($M - \mu$) values. Therefore, we can translate all the statements in Step 1 about the relative likelihood of different *M* values into statements about the relative likelihood of different estimation errors. So "*M* is most likely to fall at, or very near, μ" translates to "($M - \mu$) is most likely to be zero, or very small." The height of the curve tells us how progressively larger values of ($M - \mu$), which occur when *M* falls progressively farther to the right of μ, are progressively less likely. The likelihood that ($M - \mu$) = MOE is only about one-seventh as great as the likelihood that ($M - \mu$) = 0, so estimation errors as large as MOE are relatively rare. The curve keeps decreasing beyond the shaded area, so ($M - \mu$) values greater than MOE do occur—and would give red CIs in ESCI—but are progressively even less likely as ($M - \mu$) increases further. [All those statements refer to ($M - \mu$) being positive, meaning *M* falls to the right of μ, but the curve is symmetric, so we can make similar statements about means that fall to the left of μ, for which ($M - \mu$) is negative.] In brief, the curve tells us that estimation errors near zero are most likely, and illustrates how larger estimation errors become progressively less likely.

3. This is the crucial step, probably deserving a drum roll. In practice we don't know μ, and we have only the single value of *M* from our sample. But all the statements in Step 2 about estimation error apply. Therefore, we know that ($M - \mu$) close to zero is most likely, meaning our *M* has most likely fallen close to the unknown μ. Larger estimation errors are progressively less likely, and, correspondingly, it's progressively less likely that our *M* has fallen those larger distances from μ. Now the drum roll: We can take the curve at the bottom of Figure 4.4, which is centered on μ, and center it instead on our *M*—it will indicate the relative likelihood of

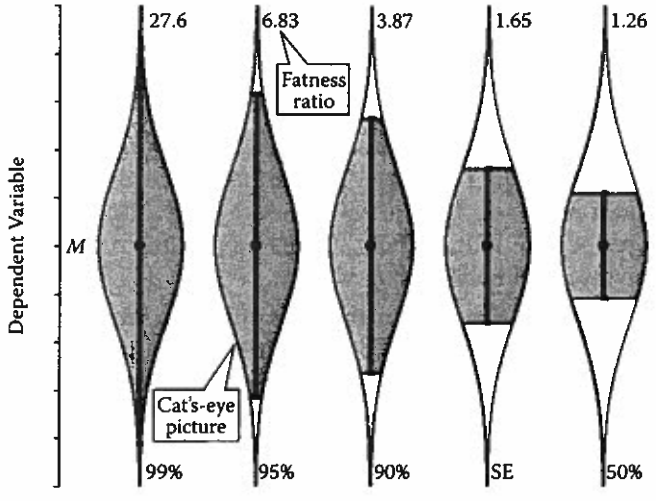


FIGURE 4.5
Cat's-eye pictures of CIs for several levels of confidence, and of SE bars. For each interval the sampling distribution of estimation errors, which is the curve at the bottom in Figure 4.4, plus its mirror image is centered at *M*, the sample mean. In each case the area between the curves that corresponds to the extent of the bars is shaded. The numbers at the top are the fatness ratio, which is the greatest fatness of the shaded area (at *M*) divided by the fatness at either end of the interval.

all possible sizes of estimation error of our *M*. Taking that curve and its mirror image and centering them on *M* is what I did to create Figure 4.5.

4. Figure 4.5 shows several examples of what I call a *cat's-eye picture*, or simply a *cat's eye*. The cat's eye comprises the sampling distribution of estimation errors and its mirror image, centered on *M*, in the middle of the interval. The *fatness* of the picture, meaning the horizontal width between the two curves, indicates the relative likelihood of different estimation errors, within and beyond the CI. Small estimation errors are most likely, as signaled by the fattest part of the picture near *M*. In other words, our *M* has most likely fallen close to μ. Therefore, values close to *M* are the most plausible for μ and are our best bets for μ. The cat's-eye picture then gets progressively less fat for values toward either end of the CI, reflecting the fact that larger estimation errors are progressively less likely, and therefore values

The *cat's-eye picture* is my name for any of the representations of intervals in Figure 4.5 complete with two sampling distribution curves and a shaded area.

Fatness is my term for the horizontal width of the cat's-eye picture, as in Figure 4.5. Fatness is greatest at *M* and decreases smoothly for values that are progressively farther from *M*. Fatness indicates how the plausibility for μ varies for values within and beyond the CI.

farther from M are progressively less plausible for μ. Values in the unshaded tails beyond the CI are progressively even less fat, and therefore even less plausible for μ, although not impossible. A statistician may prefer to say the fatness of the cat's-eye picture shows how the *relative likelihood* of various values for μ varies within and beyond the CI. Note that fatness, plausibility, and relative likelihood decrease smoothly with distance from M. There's no sudden jump at either limit of the CI. There's little difference in the plausibility of values just inside or just outside a CI.

In brief, the cat's eye summarizes the distribution of plausibility that your estimation error is small, medium, or large. It's highly revealing about what intervals are telling us, so I see it as a beautiful picture.

The shaded area is the region between the two curves that corresponds to a particular CI, as in Figure 4.5. For the 95% CI, the shaded area is about seven times as fat at M as it is at either end of the CI. This reflects the fact that, in the dance of the means, a 95% CI is more likely to land so its M is very close to μ, than it is to land so its upper limit (UL) is very close to μ. In fact, about seven times more likely. And the same for its LL. We could refer to that ratio as the *fatness ratio*. In other words, the fatness ratio is the fatness, or width, of the shaded area at M, divided by the fatness at either the UL or LL of the CI. Figure 4.5 reports near the top the fatness ratios for the different intervals. It's 6.83 for a 95% CI. In other words, M is about seven times as plausible, or seven times as good a bet for μ, as the UL. And the same for the LL.

The *fatness ratio* is my name for the fatness at M divided by the fatness at either limit of a CI.

In Figure 4.5 each picture is based on the same sample, so each has the same M and s. The two curves of the cat's eye are the same for each and describe how plausibility varies smoothly over the full range of the dependent variable. The five pictures differ only in the percentage of the area between the two curves that's shaded, and that percentage equals the level of confidence. For SE bars, about 68% of the area is shaded, because SE bars mark, approximately, the 68% CI.

In Figures 4.4 and 4.5 the sampling distribution curves are normal distributions because so far I've been assuming σ is known. If we drop that assumption, the sampling distribution curves are t distributions, with $(N - 1)$ degrees of freedom. The cat's eye therefore comprises two t distribution curves rather than two normal curves. However, in most cases the shapes and ratios of fatness reported in Figure 4.5 would change only a little. For most practical purposes Figure 4.5 provides a sufficiently accurate guide, especially considering I'm proposing the cat's-eye picture to assist understanding rather than as a basis for precise calculations. For small N, however, regarding SE bars as a 68% CI can be quite inaccurate.

CI Interpretation 4: The Cat's-Eye Picture

The cat's-eye picture gives us our fourth way to interpret CIs. Our second way, in Chapter 3, referred to interpretation of the interval as giving a range of plausible values for μ. Now we can take that idea further and have in mind the cat's-eye picture that signals how plausibility for μ varies across and beyond the interval. Values close to *M* are most plausible for μ, and the cat's eye shows how plausibility, or relative likelihood, drops toward LL and UL, then decreases further beyond the interval.

In Chapter 3 we explored how, for a given sample, changing the level of confidence, *C*, requires intervals of different length. Higher *C* requires longer intervals: The 99% CI is longer than the 95% CI. There's no change in the sampling distribution of *M* at the bottom of Figure 4.4, but the shaded area extends farther, to include 99% rather than 95% of the total area under the distribution curve. In Figure 4.5, the shaded area for the 99% CI has to extend farther into the skinny tails of the curves to achieve such high confidence, and so there's a large change in fatness across the CI—Figure 4.5 tells us that the fatness ratio is about 28. For the 90% CI, the cat's-eye shading doesn't extend as far. For 50% CIs, only the fat center of the picture is shaded, and so there's little variation in plausibility within the interval, and the fatness ratio is only a little greater than 1. For SE bars there's also only small variation in plausibility within the bars, but quite large tail areas beyond them. Note that the 50% CI is about one-third the length of the 95% CI, so about half the "weight" of a 95% CI is concentrated in the middle third of its length, where fatness varies little.

Interpretation 4 of a CI. The cat's-eye picture describes how the plausibility, or relative likelihood, that a value is μ is greatest at *M*, in the center of the CI, then decreases smoothly to either end of the CI, then drops further beyond the interval.

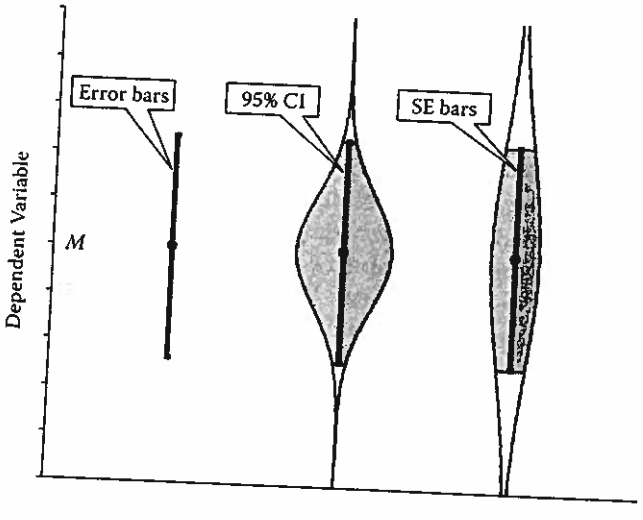
I'd like to insert an example here, in which a published researcher has referred to the relative plausibility of points within a CI. Alas, I haven't found one, so if you choose to use the cat's-eye picture to help you interpret a CI calculated from your data, you'll be at the forefront of CI interpretation. I'll now make a few suggestions of how we could use the cat's-eye picture to interpret a CI, based on the first sample in Figure 4.4. That's the sample at the bottom, just above the curve, which has $M = 54.3$. The MOE is 10.1 and so the 95% CI is [44.2, 64.4]. Thinking about cat's-eye pictures could lead you to note that

- A value of 54 is about seven times as plausible for μ as a value of 44, or one of 64;
- You can be 95% confident μ lies between about 44 and 64, and 50% confident it lies between about 51 and 57.5—which is about the middle third of the 95% CI (and approximately the 50% CI); and
- Plausibility doesn't change much over the interval from 51 to 57.5, but drops outside that interval.

All those observations are justified and could guide your interpretive comments, but they should not be taken as exact probability statements about our particular interval. As always, bear in mind the dance of the CIs, and remember: Our interval just might be red.

Here's another way to think about the message of Figure 4.5. The distribution of estimation errors, as shown by the curve at the bottom of Figure 4.4, conveys the full information in a sample about μ. Any CI tells us about that full distribution, but to interpret it correctly we need to pay careful attention to the level of confidence, C. It's C that tells us what proportion of the distribution the CI reports, and what percentage of the cat's-eye picture is shaded. I recommend routinely using 95% CIs so we can become skilled at interpreting the 95% cat's eye, and don't need to worry about intervals with the various other shaded shapes shown in Figure 4.5.

I'm not suggesting that journals should publish cat's-eye pictures whenever they show CIs, but I do suggest that imagining a cat's eye can help understanding and interpretation. In Figure 4.6 the error bars on the left are undefined: As in Figure 4.1 we've not been told what they represent, so we can't interpret them and, in particular, can't imagine the cat's eye. If the bars show a 95% CI, the cat's-eye picture in the center is correct, but f they show SE bars, the cat's eye on the right is correct. Those two interpretations of the bars must come from different data sets, with the SE bars signaling much less precise estimation than the 95% CI achieves: For the CI, the plausibility is more heavily concentrated around *M*. Therefore, for he SE bars, the *N* of the data set is smaller, and/or *s* is larger than for the 5% CI. Contrast with Figure 4.5 where all the figures come from the same



JRE 4.6
he left are undefined error bars. In the center is the cat's-eye picture if the error bars esent a 95% CI. On the right is the cat's-eye picture if they are SE bars.

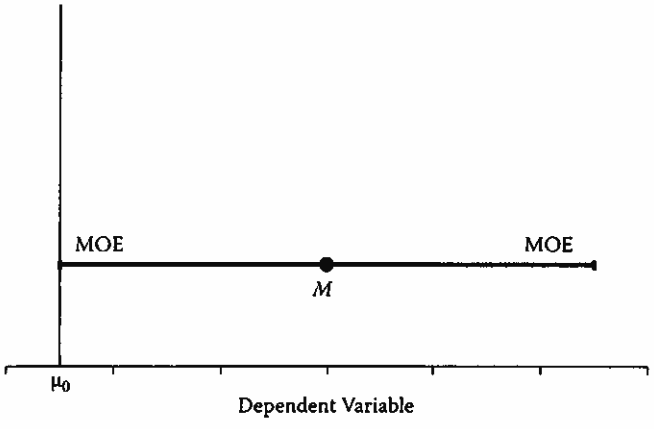
data. Now check that the fatness ratios of the two shaded areas are as you would expect, and as shown for the 95% CI and SE bars in Figure 4.5. Check that the general shapes of the shaded areas match those shown in Figure 4.5. The two cat's eyes in Figure 4.6 emphasize the dramatic difference between SE bars and 95% CIs, and reinforce my conclusion that it's a tragedy the two are often confused, and that we're often not even told which we are seeing. Can you look at any error bars, as on the left in Figure 4.6, and imagine the appropriate cat's-eye picture superimposed on them?

Finally, I should mention again our first and most basic way to interpret CIs. Whenever considering a single CI calculated from data, bear in mind the infinite dance from which it came. Here you can think of that as an infinite dance of estimation errors, many being small, some medium, and just a few large. The cat's-eye picture summarizes the distribution of plausibility that your estimation error is small, medium, or large. I find the cat's eye highly revealing about what intervals are telling us, so I see it as a beautiful picture—especially the 95% CI cat's eye. I hope you can share this feeling of beauty.

Confidence Intervals and *p* Values

I hesitate to mention *p* values, but they give us the fifth way to interpret CIs. It's my least favorite approach, and in Chapter 1 I reported evidence that CI interpretation is better if NHST is avoided. Even so, it's worth discussing the link between CIs and *p* to give a more complete picture. It's also valuable to be able to read a *p* value and generate in your mind's eye the corresponding CI. In psychology, almost all statistics textbooks explain *p* values first, then may or may not cover CIs. In some other disciplines most textbooks explain CIs first, then NHST and *p* values. Research is needed on the extent to which order might influence the quality of learning and number of misconceptions. I suspect CIs first may be better. Which order did your first statistics textbook use? What's your opinion about order?

In Chapter 1, I introduced the rule that if a 95% CI includes μ_0, we can't reject H_0. Therefore, two-tailed $p > .05$. If the interval does *not* include μ_0, we reject H_0 and note that $p < .05$. This rule makes sense because if μ_0 lies outside the interval it's a relatively implausible value for μ, and therefore it's reasonable to reject it. Conversely, if μ_0 lies in the interval, it's a plausible value for μ, and so we can hardly reject the hypothesis that states it is the value of μ. The boundary case occurs if a 95% CI falls so either of its limits is exactly at μ_0, as in Figure 4.7, in which case $p = .05$.



IGURE 4.7
, CI that falls so one of its limits is at the null hypothesized value, μ_0.

This rule generalizes for other levels of confidence simply by adjusting he *p* value. For a 99% CI, the *p* value is .01 if the interval falls so imit is at μ_0. For a 90% CI, *p* is .10 if either limit is at μ_0, and so on. For a C% CI with a limit at μ_0, $p = (1 - C/100)$. Figure 4.7 illustrates any of these cases, nd Box 4.2, which is an optional extra, explains that formula.

Note my slightly awkward wording, for example, "if the interval falls so ither limit is at μ_0," when it might seem clearer to say, "if μ_0 is at one of he limits of the interval." I prefer the first rather than the second wording o emphasize it's the interval that's the variable, not μ_0 or μ, but sometimes 'll use the second wording. In any case, keep in mind the dance of the CIs.

Box 4.2 explains that, if a C% CI has a limit at μ_0, then two-tailed $p = 1 - C/100$). The box makes a simple relation look pretty complicated, ut does illustrate the fact that NHST and CIs are closely linked. Indeed, hey're based on the same underlying statistical model and assumptions. Given this common theoretical base, it may be surprising that they lead to uch different thinking and consequences. Anyway, now for a picture that hope makes the relation between C and *p* easy to grasp.

he CI Function

igure 4.8 shows the CI function, another beautiful picture that reveals nore about CIs. The easiest way to understand it may be to fire up **ESCI hapters 1–4**, and go to the **CI function** page. Drag the big slider up and lown, and see the movable CI, shown in heavy black in Figure 4.8, sweep ıp and down, changing in length as it goes. The two limits of the interval nark out the big double curves that are the CI function. The left vertical axis

BOX 4.2 THE RELATION BETWEEN *C* AND *p*

I stated that if a C% CI has a limit at μ_0, then two-tailed $p = (1 - C/100)$. Here's an explanation of that relationship. To use NHST for a single group, we calculate the obtained value of t with $(N - 1)$ degrees of freedom by using the formula

$$t_{obt}(N-1) = (M - \mu_0)/(s/\sqrt{N}) \tag{4.3}$$

which implies that

$$(M - \mu_0) = t_{obt}(N-1) \times (s/\sqrt{N}) \tag{4.4}$$

After calculating $t_{obt}(N - 1)$ we'd use tables or software (such as ESCI **Normal z t**) to find the corresponding two-tailed p value, which by definition is the probability of obtaining $t > |t_{obt}(N-1)|$ if the null hypothesis H_0: $\mu = \mu_0$ is true. To put it another way, $t_{obt}(N - 1)$ is the critical value of t for that p value, and we write that critical value as $t_{(1-p)}(N - 1)$. An example is that the critical value of t for a p value of .05 is $t_{95}(N - 1)$. We can therefore substitute $t_{(1-p)}(N - 1)$ in place of $t_{obt}(N - 1)$ in Equation (4.4) to obtain

$$(M - \mu_0) = t_{(1-p)}(N-1) \times (s/\sqrt{N}) \tag{4.5}$$

Now consider CIs and recall from Chapter 3 that Equation (3.3) gave, for a 95% CI

$$\text{MOE} = t_{95}(N-1) \times (s/\sqrt{N})$$

so for a C% interval we would use

$$\text{MOE} = t_{C/100}(N-1) \times (s/\sqrt{N}) \tag{4.6}$$

Notice that the MOE of the CI in Figure 4.7 is simply $(M - \mu_0)$, so, for the CI in Figure 4.7, Equation (4.6) gives us

$$(M - \mu_0) = t_{C/100}(N-1) \times (s/\sqrt{N}). \tag{4.7}$$

Compare Equations (4.5) and (4.7) to find that

$$t_{(1-p)}(N-1) = t_{C/100}(N-1). \tag{4.8}$$

Therefore, $(1 - p) = C/100$ or, equivalently, for the situation of Figure 4.7 we have $p = (1 - C/100)$, which is the relationship we wanted to explain.

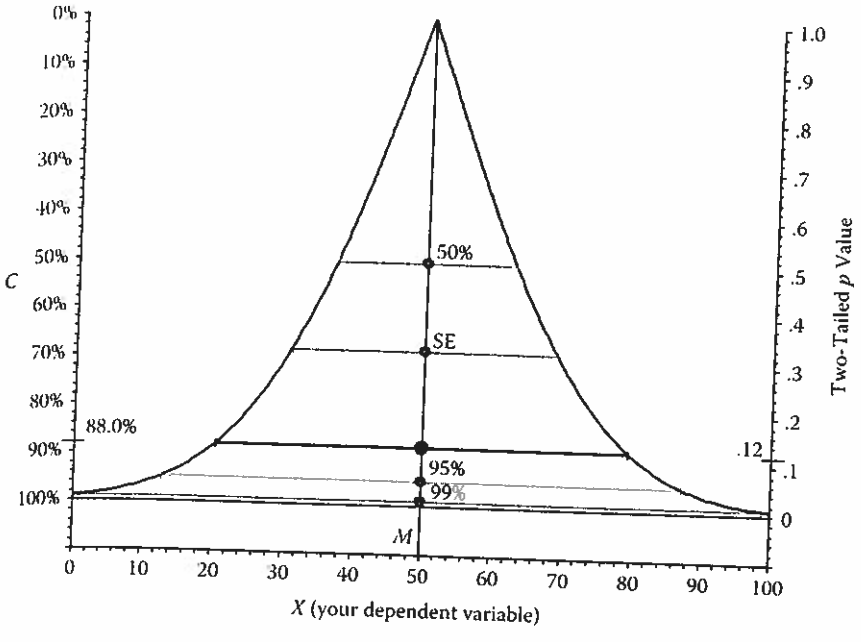


FIGURE 4.8
The CI function, from the **CI function** page of **ESCI chapters 1–4**. It plots the level of confidence C (left axis) and two-tailed p value (right axis) against the lower and upper limits of a CI. Fixed intervals are shown in gray, for comparison, and the movable interval, currently set to be an 88% CI, is shown in heavy black.

shows the level of confidence C of the movable CI, and the right vertical axis shows the two-tailed p value for that CI when the null hypothesized value lies at either limit of the interval—as illustrated in Figure 4.7. The CI function is sometimes known as the p value function. A version was introduced by Poole (1987), and it was discussed by Rothman (2002, Chapter 6).

Can you figure out what's going on here, and why the curves sweep out so sharply at the bottom, at high values of C? Compare with Figure 4.5. For C around 90 or more, the cat's-eye picture in Figure 4.5 is skinny near either of its limits, and so large increases in CI length are needed to yield a modest increase in shaded area and C. Therefore, in Figure 4.8 the CI function sweeps out dramatically near the bottom. It all hinges on the shape of the normal or t distribution that defines the cat's eye and, to achieve high levels of confidence, our CI needs to extend into the tails of that distribution.

I'm not going to write out lots of small steps to suggest how you can use the **CI function** page in ESCI, as I did for **CIjumping** in Chapter 3. Explore the CI function as you wish. Here are a few pointers.

- Enter *M*, *SD*, and *N* for your single sample. Adjust these values to control the position and width of the function.
- Observe the relation between CI length, *C*, and *p* values. Compare with Figure 4.5.
- Click below red 3 to display a cat's-eye picture on the main CI. Watch how the shading changes as you zoom the CI up and down. The fatness ratio, which is reported below red 3, also changes, but the two mirror-image curves of the cat's eye remain the same. The percentage of the area shaded always matches the level of confidence. Use the spinner below red 3 to change the amount of bulge: This changes the vertical scale of the cat's-eye picture, but doesn't change the fatness ratio or the interpretation.
- Click below red 1 to regard *SD* as the population value, so the CI is based on z, or to regard *SD* as the sample value, in which case the CI is based on *t*. See how the function changes, as well as the shape of the cat's eye. If *N* is large there's little change, but if *N* is small there's considerable change—because, as you know, the *t* distribution differs greatly from the normal distribution at very small *df*.

Translating Between a 95% CI and *p*

The next step is to investigate how, in your mind's eye, to skip back and forth between any *p* value and the corresponding 95% CI. First, notice in Figure 4.5 or Figure 4.8 that a 99% CI is roughly one-third longer than a 95% CI. If μ_0 lies at the end of a 99% CI, *p* is .01, so if we're looking at a 95% CI, we know that if μ_0 lies about one-third of MOE beyond the end of the interval, *p* must be .01. Apply that logic for CIs with other levels of confidence, and we can read any *p* value from where our standard 95% CI lies in relation to a hypothesized value. That turns out to be very useful.

Figure 4.9 shows the left arms of several CIs, with the 95% CI in bold as a reference interval. The dotted vertical lines mark where the CIs with various levels of confidence, *C*, have their lower limit. These lines are labeled with the corresponding two-tailed *p* value, meaning *p* when μ_0 lies at the position of the dotted vertical line. So, for example, the 99% CI has its lower limit at the line labeled with the *p* value .01. The fractions indicate distances from the left end of the 95% CI, in units of MOE (of that 95% CI), so the 99% CI extends approximately an extra one-third of MOE beyond the end of the 95% CI.

I suggest it's worth remembering those four fractions, as approximate benchmarks for the corresponding four *p* values. They state that $p = .01$ when μ_0 lies about one-third of MOE beyond the end of a 95% CI, and $p = .001$ when it lies about two-thirds MOE beyond. Inside a 95% CI, about

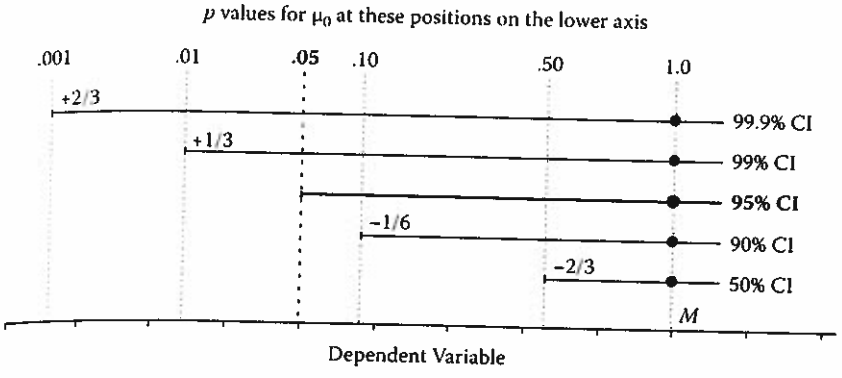


FIGURE 4.9
Left arms of a number of CIs for the same data, as labeled at the right. The null hypothesized value, μ_0, can be given any value on the horizontal axis at the bottom. If μ_0 has the value marked by the heavy dotted vertical line, then $p = .05$, as marked at the top of that line. If μ_0 has the value marked by any of the light dotted vertical lines, then two-tailed p has the value marked at the top of that line. The dotted vertical lines also mark the ends of the CIs corresponding to the various p values. The four fractions are approximate benchmarks worth remembering. They state how much longer or shorter the MOEs are than the MOE for the 95% CI, which is shown in bold and serves as a reference.

one-sixth of MOE back from the end gives $p = .10$, and about two-thirds of MOE back from the end (or one-third MOE out from M) gives $p = .50$. Keep those benchmarks in mind, and interpolate for values between those reference points.

For practice, look back at Figure 1.1 and eyeball the p values from the figure showing the 95% CIs for Lucky and Noluck. For Lucky, zero (the null hypothesized value) lies beyond the interval but not as far away as one-third of MOE. So p will be less than .05 but not as small as .01. If you estimated around .02, you are getting the idea fast. For Noluck, zero is further back from the limit of the CI than one-sixth MOE, so we know p is greater than .10, but zero is well beyond the benchmark for .50. If you estimated around .20, you are, again, doing very well.

Here's a bigger challenge: Run this guesstimating backwards. Read the first presentation of Lucky–Noluck, which reported M and p, and try to generate in your mind's eye the figure showing the two 95% CIs. Note that we're using $\mu_0 = 0$, and for each study you know M and the p value. For Lucky, M is around 3.6 and p is .02, so you know the 95% CI will extend most of the way from 3.6 toward zero, but will stop short by less than one-third of MOE. Therefore, the lower limit of the 95% could be around 0.5 or a little more, which suggests that MOE is around 3. Therefore, the 95% CI is roughly [0.6, 6.6]—which Figure 1.1 shows is a pretty good eyeballing result. For Noluck, consider M of about 2.2 and p of about .20, so we know the interval extends from 2.2 past zero, and by more than one-sixth MOE

because p is greater than .10. I guesstimated MOE to be a bit more than 3, which Figure 1.1 shows to be about right, or maybe a bit short. It's close enough for eyeballing purposes.

You can play around with p and the position of μ_0 in relation to a 95% CI by using the **CI and p** page of **ESCI chapters 1–4**. Use the big vertical slider to move the 95% CI up and down, changing its position relative to μ_0, which remains fixed. Click to show or hide an axis showing a wide range of p values, the four benchmarks, and the accurate p value. Use the page for guessing games: For example, hide p values and the benchmarks, position the CI as you wish, and then compete with someone else to estimate p. Click to reveal the p value. Who's more accurate? Another example: Compete with someone else to position the CI to give some stated target p value. You'll both quickly become fast and accurate, and you'll have a useful skill that most researchers lack—or perhaps don't even realize is possible.

In a world full of p values, but lacking CIs, it can often be revealing to generate in your mind's eye what the 95% CIs would look like, given only some p values. Note, for example, what happens for $p = .30$, or some other value that's clearly not statistically significant. The CI is quite long relative to the ES—which is the distance from M to μ_0—so there's considerable uncertainty and, almost certainly, no justification for accepting the null hypothesis. Translating to a CI may be the best way to interpret a p value.

I've mostly been using the normal distribution, to keep things simple. I've therefore usually been assuming σ is known, or that we're using very large samples. I mentioned that if we drop that assumption and use the t distribution with $df = (N - 1)$, the cat's-eye pictures would change shape a bit, and the CI function would be different—most noticeably when N is very small. The **CI and p** page allows you to click to display the 95% CI and p values based on t rather than z: The two are displayed side-by-side, so you can compare, consider the benchmarks, and see how the difference varies as you change N. The figures, from Figure 4.4 onward, and the benchmarks I've suggested describe intervals and p values based on z. For our estimation purposes, they are all accurate enough to be practically useful. Just bear in mind that if N is small, say, less than about 10, our eyeballing may be a little astray.

I discussed cat's-eye pictures, benchmarks for p, and various other things in this chapter in an article in the journal *Teaching Statistics* (Cumming, 2007).

CI Interpretation 5: The Relation Between CIs and *p* Values

The fifth way to interpret a CI is in terms of p values: Note where the 95% CI lies in relation to μ_0, then estimate p. It's my least favorite approach to finding meaning in a CI, but it's still worth being able to do such eyeballing. More useful may be the ability to run this process backwards and generate in your mind's eye the 95% CI, given only μ_0, the sample

mean, and *p*. If you can do that, you have additional insight into the great mass of research that's published with *p* values, but not CIs. And you have a valuable skill I suspect is rare.

Interpretation 5 of a CI. Where a 95% CI falls in relation to the null hypothesized value signals the *p* value. Use the benchmarks to eyeball an estimate of *p*.

At this point you may be expecting examples of how CIs can be interpreted with reference to *p* values. I'm not going to include any, however, because I don't want to encourage this approach. You'll recognize such examples easily enough by mention of inclusion or exclusion of a null hypothesized value, or stating of a *p* value. There are also exercises at the end of the chapter.

One-Tailed Tests, One-Sided CIs

So far in this book I've discussed two-tailed NHST and two-sided CIs. You may have wondered whether I've avoided one-tailed tests because they are a strength of NHST that estimation can't match. Not so! *One-sided CIs* are analogous to one-tailed tests but, as usual, the estimation approach is better. Strangely, most NHST textbooks describe one-tailed tests, but even textbooks that include CIs rarely mention one-sided intervals—which can, however, be useful, and are worth knowing about. I'll use an example and Figure 4.10 to explain.

One-sided CIs are analogous to one-tailed tests, but more informative.

Suppose a well-established therapy for a certain type of wrist injury in elite cyclists gives, after 20 sessions, the ability to lift with the injured wrist an average 520 g in a standard exercise device; the SD is 88 g. You are estimating the effectiveness of a new therapy, but are only interested if it's better. You observe a group of $N = 31$ cyclists use the new therapy and calculate $M = 548$ g for their ability to lift in the exercise device.

Let's assume that the mean and SD stated for the standard therapy are population values, perhaps estimated precisely by a meta-analysis. So, for NHST, we'll use H_0: $\mu_0 = 520$ g as the null hypothesis. Then Figure 4.10 panel A shows the sampling distribution of M, for $N = 31$ and $\sigma = 88$ g, and the rejection region for a two-tailed statistical significance test with $\alpha = .05$. We'll reject the null hypothesis if M falls in a shaded tail area, meaning $M > 551$ or $M < 489$. However, we're only interested in whether the new therapy is *better*, so we could justifiably use a one-tailed test, as illustrated in panel B. That rejects the null hypothesis if $M > 546$—an easier criterion to meet than the 551 of the two-tailed test. The key thing about the one-tailed test is that, yes, it makes statistical significance easier to obtain, but we should only choose it when the situation is genuinely

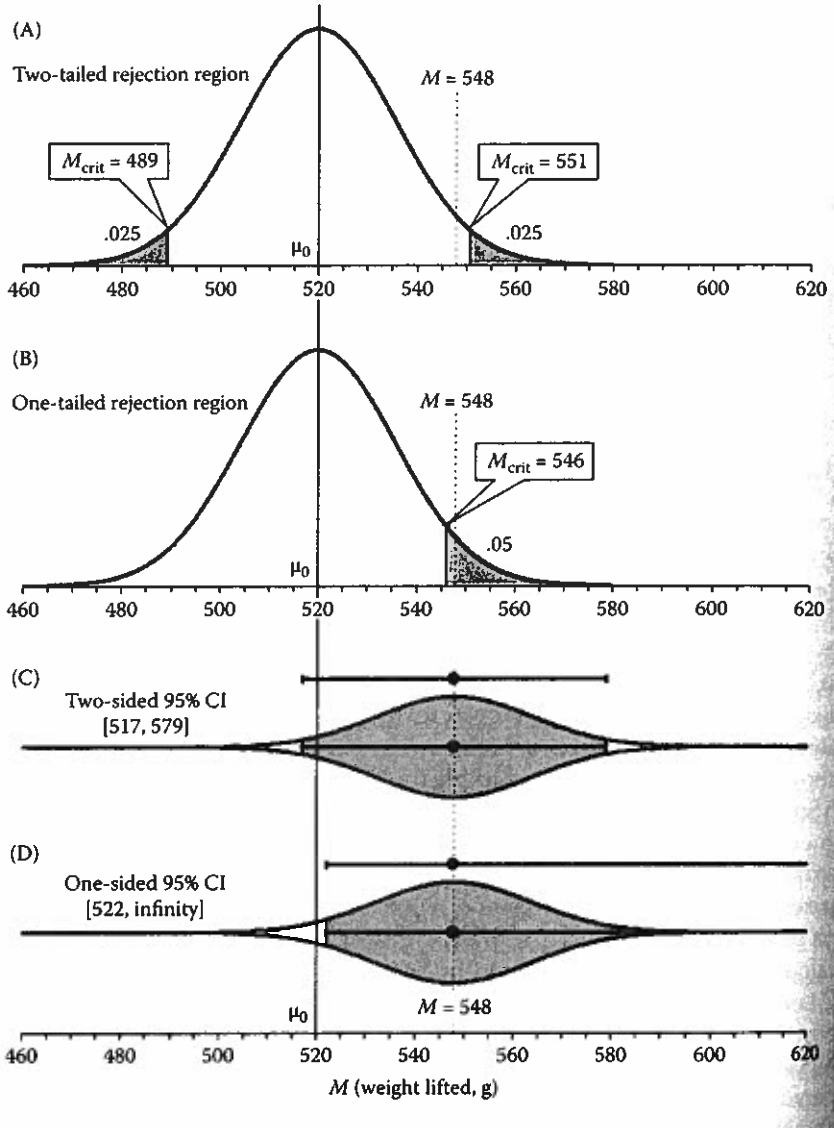


FIGURE 4.10
Panel A shows the sampling distribution of M, the mean weight lifted by a group of N = 31 cyclists recovering from a wrist injury, and the two-tailed rejection region when $\alpha = .05$ and $\mu_0 = 520$ g is the population mean and $\sigma = 88$ g is the population SD. Panel B shows the same for a one-tailed test. Critical values are shown as M_{crit}. Assuming $M = 548$ is observed, as marked by the dotted vertical line, panel C shows two images of the two-sided 95% CI and panel D a one-sided 95% CI.

asymmetric and we only care about one specified direction of effect. Also, we have to commit in advance to using the one-tailed test.

For your $N = 31$ cyclists you calculated $M = 548$ g, as marked by the dotted vertical line in the figure. That would reject the null hypothesis, with $\alpha = .05$, if we had committed to a one-tailed test, but not if we were using the two-tailed test. Panel C shows the 95% CI around $M = 548$. That conventional two-sided interval includes the null value, 520, consistent with the result of the two-tailed test. Panel D shows the 95% one-sided CI around $M = 548$ to be [522, infinity], meaning the interval extends indefinitely to the right. The one-sided CI does not include 520, consistent with the result of the one-tailed test. If we think of whether or not the CIs include the null value, there's a direct correspondence between, respectively, the two- and one-tailed tests, and the two- and one-sided intervals.

Note that the curves of the cat's-eye picture are identical in panels C and D. In Figure 4.5, CIs with various levels of confidence have different percentages of the cat's eye shaded, but the curves are always the same. This is similar in Figure 4.10 for two-sided and one-sided CIs: The difference is in which 95% of the area we choose to shade. For the two-sided CI, 2.5% of the area of the cat's eye lies beyond each end of the CI, whereas for the one-sided CI, all 5% of the area outside the interval lies to the left. The one-sided CI extends indefinitely to the right, because the cat's-eye picture extends indefinitely both ways. In practice there's an upper limit to what any human wrist can lift, but I can't give an exact value for that, so I'll state "infinity" as the upper limit (UL) of the CI.

I should mention that we need to be careful of how a one-sided CI is described. I've seen an interval such as that shown in panel D described as

- A lower end-point CI
- A lower one-sided CI
- An upper one-sided CI
- An upper-tailed one-sided CI

suggest that "lower end-point CI" is the ast ambiguous, and thus what we might refer. However, always state the numeri- l limits or refer to a figure to make clear which one-sided interval you ean—is the short arm below or above the mean?

Refer to a one-sided CI like that in Figure 4.10 as a *lower end-point CI*, but also report values or a figure for clarity.

We can use any of our approaches to interpreting a CI to think about one-sided CI. Considering the dance of the CIs, in the long run 95% of ie-sided CIs will include the population mean, just as we expect for two- led intervals. In the two-sided case, the red CIs that don't capture the pulation mean will split between those missing high and those missing

low, whereas all 5% of one-sided intervals that miss will miss high (for the lower end-point case, as in panel D). We can say we're 95% confident our one-sided interval includes the true value. We can say the lower limit (LL) of the one-sided CI (522 for our example) is a likely lower bound for the true value, meaning that for 5% of replications the LL will exceed the true value. Compare that with the LL of the two-sided CI (517 for our data), which is also a likely lower bound, but with a different meaning of "likely" because the LL of the two-sided CI will exceed the true value for just 2.5% of replications.

Calculating a One-Sided CI

ESCI doesn't display one-sided CIs, but it's easy to calculate them if you wish. Note again the cat's-eye picture in panel D, in which 5% of the area is below the interval. The 90% two-sided CI has 5% of the area of the cat's eye below and another 5% above, so its LL is the same as the left limit of the one-sided 95% CI illustrated. To find a 95% one-sided interval, use ESCI or any other software to find the 90% two-sided interval, then choose the LL or UL of that CI as your single limit of the one-sided 95% CI that you seek.

The short arm of a one-sided 95% CI is the same as either arm of a two-sided 90% CI.

You may recall $z_{95} = 1.960$ is the critical value we use to calculate the MOE of a two-sided 95% CI. For a two-sided 90% CI we use $z_{90} = 1.645$, so that's the value we need to calculate the one-sided 95% CI. Now, 1.645 is 16% less than 1.960, so the lower arm of the one-sided CI in panel D should be 16% (about one-sixth) shorter than either arm of the two-sided CI in panel C. To my eye, that's about what the figure shows. Here's a different example: If $M = 10.0$, 95% CI [6.1, 13.9], then the upper end-point one-sided 95% CI is [–infinity, 13.3].

Two-Sided or One-Sided CIs?

It's useful to know about one-sided CIs because they provide an additional option. A one-sided CI parallels the one-tailed test and gives additional information beyond the test result, but I'd rather think about one- and two-sided CIs without reference to NHST. The key is to bear in mind panels C and D of Figure 4.10. Think of the cat's-eye picture and decide which CI is more appropriate, given your research questions. Arguably you should do that in advance of collecting data, just as you need to commit to a one-tailed test in advance. However, if you refrain from interpreting CIs merely to carry out NHST, and appreciate how one- and two-sided CIs relate, I'm comfortable with your choosing between a one- or two-sided CI as you analyze your data. Is it more informative for your readers who are interested in the new wrist therapy if you report and discuss your

findings using the two-sided CI in panel C or the one-sided CI in panel D? Which fits better with your research aims?

Examples 3.2 in Chapter 3 included a CI for the vaccination rate needed to avoid an epidemic, and a CI for the tiny distance down to which a prediction of Einstein's special theory of relativity has been confirmed. In each case the researchers chose to interpret the UL of their two-sided CI, that being in each case a conservative value. Each team of researchers could reasonably have chosen instead to use upper end-point one-sided CIs.

I don't use one-sided CIs often, but it's unfortunate that they are usually ignored even by textbooks that cover CIs. They provide a useful additional option for understanding and communicating research results. Also, I suspect that understanding one-sided CIs, via the cat's-eye picture, probably increases our understanding of estimation in general.

In the next chapter I discuss replication, and what's likely to happen if you repeat your experiment over and over. That discussion follows on from what we've been considering in this section and gives the sixth approach to CI interpretation.

It's time for you to write your take-home messages from this long chapter. Have you been jotting them down as we go? Here are some hints, before I give you my list:

- Many types of error bars. Inferential and descriptive statistical information.
- Beautiful pictures of CIs. The fourth way to interpret CIs.
- The relation between CIs and p values. The fifth way to interpret CIs.

xercises

4.1 You are listening to a research talk about an evaluation of a children's fitness program. The speaker displays Figure 4.11, which depicts the average improvement in performance scores after the program. You raise your hand and ask what the error bars represent. The speaker is flustered and anxiously consults colleagues, but eventually says "a confidence interval." You decide to wait until after the talk to seek confirmation it's a 95% CI. Assuming it is, suggest three interpretations without mentioning NHST or a p value. In each case identify which interpretation you're using.

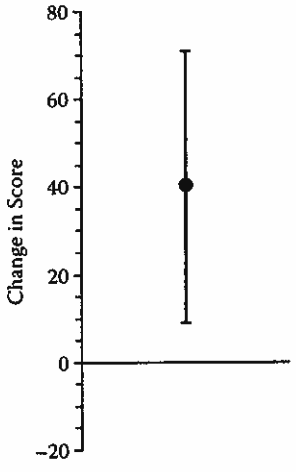


FIGURE 4.11
Mean improvement in children's performance after a fitness program.

4.2 The speaker is unsettled by your question, stops and consults notes, then apologizes and states that the figure shows SE bars. On this assumption suggest three interpretations, without mentioning NHST or a *p* value.

4.3 Now interpret the result using a *p* value, first assuming a 95% CI, then assuming SE bars.

4.4 Suppose the speaker now stated that $N = 5$. Would any of your previous answers change? How?

4.5 Use the **Normal z t** page of **ESCI chapters 1–4** to find the fatness ratio for the cat's-eye picture for a 95% CI when $N = 30$ and when $N = 5$. Compare with the ratio for σ known. *Hint*: You can use **Normal z t** to find the height of the *z* or *t* distribution at any point by clicking near red 5 to turn on **Heights**. Click to display **Two tails** then move the slider to find the height you want. To find, for example, the fatness ratio for a 95% CI, you could divide the height at the center of the distribution by the height at the .05 tail boundary.

4.6 Find an example in your own work or in a journal article, or invent an example, for which using the cat's-eye picture is helpful for interpreting a CI. Explain.

4.7 Use the **CI and p** page to find a benchmark, additional to those shown in Figure 4.9, for $p = .20$. *Hint*: Some eyeballing is required.

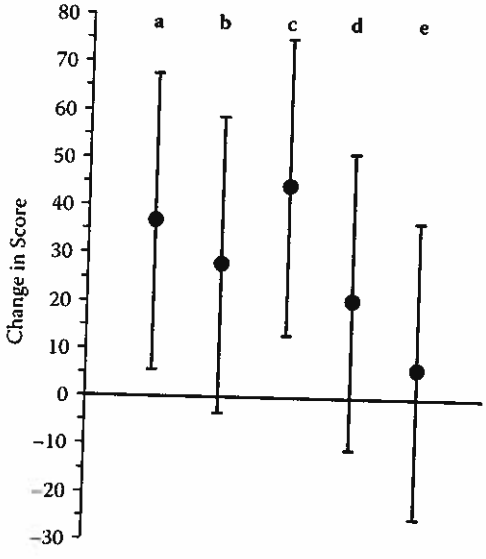


FIGURE 4.12
Some example means with 95% CIs.

4.8 Suppose Figure 4.12 shows means and 95% CIs, and that N is large or σ known. The vertical axis represents change scores and so we choose zero to be the null hypothesized value. Considering each result independently, estimate two-tailed p for each result.

4.9 Sketch a figure showing the mean and 95% CIs for the following results. Consider each independently, and assume in every case that N is large or σ known, and zero is the null hypothesized value:

(i) $M = 5$ and two-tailed $p = .60$

(ii) $M = 20$ and two-tailed $p = .15$

(iii) $M = 15$ and two-tailed $p = .002$

4.10 In Figure 4.12, suppose result **b** is a 90% CI. Estimate p. Suppose result **d** is a 99% CI. Estimate p.

4.11 Find, in your own discipline, a report of a one-sided CI, or (probably easier) a one-tailed test. Is it used appropriately? What would you recommend?

4.12 Before reading on, check your own take-home messages. Revise or expand if you wish.

Take-Home Messages

- In a figure, a dot may be preferable to a column to represent a sample mean or other point estimate, because the sharp top of a column may suggest unwarranted precision.
- It's often valuable to show CIs as error bars in a figure. CIs provide inferential information, meaning information based on the sample that informs us about the population.
- The familiar error bar graphic is, unfortunately, used to show a range of different quantities. It's essential that every figure with error bars states clearly what the bars represent.
- Unfortunately, SE bars are common in some disciplines. SE bars don't give clear descriptive information or accurate inferential information. Always prefer CIs to SE bars.
- Choose 95% CIs unless there are good reasons to use a different level of confidence.
- A rule of eye is that double the length of SE bars gives, approximately, the 95% CI. The rule is reasonably accurate in many cases, but not all. For example, when $N < 10$ the 95% CI is longer than double the SE bars.
- Bars may also be used to represent descriptive information about a sample, for example, the SD or range.

- The sampling distribution of the sample mean is also the sampling distribution of estimation errors. Place that distribution (and its mirror image) on a CI and obtain the beautiful cat's-eye picture of a CI, which shows how plausibility, or relative likelihood, varies across and beyond the interval.
- *Take-home picture*: The cat's-eye picture of a 95% CI, as in Figure 4.5.
- Plausibility of a value for μ is greatest at M in the center of a CI. For a 95% CI it drops to about one-seventh at either limit. For a 50% CI it drops little to either limit but there are large tails beyond the interval. A 99% CI has to extend into the thin tails to achieve such a high level of confidence.
- The cat's-eye picture gives our fourth way to interpret CIs, by indicating how plausibility varies across and beyond the interval.
- The CI function is two smooth curves that plot the level of confidence, C, and two-tailed p value against the lower and upper limits of a CI.
- *Take-home movie*: At the **CI function** page, turn on the cat's eye and sweep the slider up and down to see how CI length and the cat's eye shaded area change with C.
- If a 95% CI lands with one limit at the null hypothesized value μ_0, two-tailed $p = .05$. Use the four benchmarks to estimate p for any position of a 95% CI in relation to μ_0. Approximately: One-third MOE of the 95% CI out from M gives $p = .50$; one-sixth MOE back from a limit of the CI gives $p = .10$; one-third MOE beyond a limit gives $p = .01$; and two-thirds MOE beyond a limit gives $p = .001$.
- Those benchmarks also allow imagining in the mind's eye the 95% CI, given μ_0, M, and p. That's a useful ability in a world that reports p values, but often not CIs.
- Our fifth way to interpret CIs is in terms of p values, although this is nonpreferred and may often not give the best interpretation.
- One-sided CIs correspond to one-tailed NHST, but are more informative. Choose one- or two-sided CIs depending on the research questions and the context.