3

Confidence Intervals

Soon we'll sound the trumpets for the arrival of CIs, but we need to explore several other ideas first. Watch out for the dance of the means and the mean heap, then the margin of error. After that the trumpets won't be far away. Here's the plan for the chapter:

- The population and a random sample
- Sampling: dance of the means and the mean heap
- Errors of estimation, and the margin of error (MOE)
- Confidence intervals at last
- Reporting CIs
- Interpreting CIs: the first three approaches

There are two parts to this chapter: The first covers the first four bullet points and the second covers the last two. The first part is my version of the story that starts with sampling and finishes with CIs. Numerous textbooks give a version of that story. The main novel aspect of my version is that it focuses on understanding the extent of sampling variability—which, unfortunately, some people often underestimate. I hope ESCI pictures and simulations can help you build accurate intuitions about sampling variability. Because those pictures and simulations are so central, in the first part of the chapter I integrate ESCI activities more closely into the discussion than in any other chapter. However, I include many figures, so I hope the discussion is useful even if you don't work with ESCI as you read.

The second part of the chapter comprises the sections Reporting CIs and Interpreting CIs. These go beyond what many other textbooks say about CIs. I don't need to integrate ESCI so closely into those sections.

Population and Samples

We need to start with a population and samples from that population. Suppose you are investigating the climate change awareness of university students in your country. You decide to use the Hot Earth Awareness Test

(HEAT), which is a well-established survey—actually, I just invented it—that asks questions about a respondent's knowledge, attitudes, and behavior in relation to climate change. You would like to know the mean HEAT score for students in your country. You plan to test a sample of students to estimate that mean.

Now we do some statistical assuming. (Box 3.1 gives extra detail, but it's an optional extra.) Suppose there's a large population of students in your country, and their HEAT scores are normally distributed with mean of μ and SD of σ (Greek sigma). You take a random sample of N students from that population, obtain their scores, and calculate the mean, M, and standard deviation, s, of your sample. You'll use M as your point estimate of μ. Later you'll calculate a CI to tell us the precision of your estimate—how close M is likely to be to the unknown μ.

That statistical model based on a normally distributed population and random sampling underlies many of the most commonly used statistical techniques. It's the basis for the conventional CIs we discuss here, as well as

BOX 3.1 A STATISTICAL MODEL

A statistics textbook says something like, "Consider random variable X with distribution $N(\mu, \sigma^2)$. Let M be the mean of a random sample, size N, of X. Then M has distribution $N(\mu, \sigma^2/N)$." Those three sentences summarize the first half of this chapter, but need some unpacking. A *random variable* is simply a variable that can take some range of values, with various probabilities. The abbreviation $N(\mu, \sigma^2)$ refers, as you probably guessed, to a normal distribution with mean μ and variance σ^2, which implies standard deviation σ.

The sentences express a very widely used statistical model, based on a normally distributed population of X scores, with mean μ and standard deviation σ. The model considers random samples of X scores, each sample having size N. The sample mean is another random variable M, and is normally distributed, with the same mean μ and smaller variance σ^2/N. The SD of M is thus $\sigma/\sqrt{N}$. You may know that the distribution of M is referred to as the *sampling distribution* of M, the sample mean, and that the SD of this distribution is called the *standard error* (SE). Therefore, SE = $\sigma/\sqrt{N}$. Sample mean M is a random variable because every time you take another sample you'll get a slightly different value for M. Take numerous samples then the numerous values of M form the sampling distribution of M. The textbook probably also explains how sample statistics M and s are used as estimates of population parameters μ and σ.

for conventional NHST. It's far from the only possibility, but it's the model I use throughout this book. It makes important assumptions, notably

- *Normality.* In many cases in practice this strong assumption about the population distribution may be justified, in some cases a transformation of the dependent variable improves its appropriateness, and in some cases it's not justified and some other approach should be taken.
- *Random sampling.* There are two vital aspects: First, every member of the population must have an equal probability of being sampled, and second, all sample values must be chosen independently.

You should always keep these assumptions in mind and judge how closely they are met in a particular situation. In our example, the dependent variable is the HEAT score, which we'll refer to as *X*. It may be reasonable to assume at least approximate normality of HEAT scores in the population of all university students in your country. Considering random sampling, it's unlikely you can ensure that every student has an equal chance of being included in your sample. Perhaps you can sample randomly from a range of disciplines in a variety of universities? You'll need to judge how well you think your sampling strategy will give a sample representative of the population in ways that are relevant for your HEAT research question.

Independence, the second aspect of random sampling I mentioned, is crucially important. You need to make sure you choose each student in the sample separately, rather than, for example, choosing clusters of students who are mutual friends, or who are in the same class. As so often in statistics, care and judgment are needed. This book is not primarily about research design, so I won't extend this discussion. However, I must emphasize that the new statistics require attention to assumptions just as does NHST.

Note carefully the distinction between population and sample:

- The *population* is a supposedly infinite collection of university students in your country, or rather their HEAT scores. It's common, if slightly confusing, to talk interchangeably of the population comprising the students, or the HEAT scores. Anyway, it's the HEAT scores, our dependent variable *X*, that we assume to be normally distributed. The *population parameters* are the mean μ and standard deviation σ of the population distribution of *X* scores. The values of μ and σ are fixed but unknown—because we can't ever know the HEAT scores for every student in your country.

The *population parameters* μ and σ have values that are fixed but unknown.

- By contrast, we know the N values of X that make up our *sample*, and can calculate the obtained sample mean M and standard deviation s. The *sample statistics* M and s have particular values for a particular sample, but if we repeat the experiment—by taking another, independent sample—we would get different values for M and s.

The *sample statistics* M and s are c
lated from our sample data. They v
be different for a different sample. W
them as point estimates of μ and σ.

In other words, we don't know μ and σ, but we want to. We do know M and s for our sample, but we don't especially care about those particular values, except to the extent they tell us something useful about μ and σ.

I'm about to turn to ESCI but, as I've said, I hope you can follow the discussion whether or not you use the software. I suggest you read the exercises, as well as the main text, but skim over references to the fine details of ESCI if you wish. Many of the exercises ask questions. Whether or not you use ESCI to find answers, you can consult the section near the back of the book that provides suggested answers. In any case, focus on the statistical ideas and the many figures I've included in the book.

Exercises

3.1 Open the **CIjumping** page of **ESCI chapters 1–4**. Consult Appendix A for hints, especially the section Strategy for Getting Started With a New ESCI Page.

3.2 Figure 3.1 shows the population, which for us is an idealized representation of all the HEAT scores of students in your country. It has a normal distribution, and the figure shows it's a symmetric bell-shaped curve. Click near red 2 to display the population curve. Use the sliders to change population μ and σ, then set them back to the values $\mu = 50$ and $\sigma = 20$, which we'll assume are the population values for your country. (As you change σ, you can see the vertical scale automatically rescaling, so the curve is always displayed with a convenient vertical height.)

3.3 Click near red 2 to fill under the curve with random little blue circles, or data points, as shown in Figure 3.1. ESCI can't display the infinite number of dots that, notionally, make the population, but you get the idea.

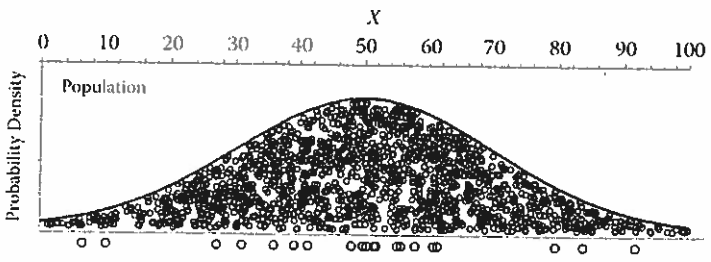


FIGURE 3.1
A part screen image from the **CIjumping** page of **ESCI chapters 1–4**. It shows the normally distributed population, with $\mu = 50$ and $\sigma = 20$, and below, a scatter of points that's a single random sample, $N = 20$, of HEAT scores taken from the population. We use X to refer to the HEAT scores.

3.4 Click the **Clear** button near red 1—nothing happens, but we're getting set. Click near red 4 to **Display data points**. Yes, still nothing, but now a dramatic moment: We're about to take our first random sample.

Text that looks like **Display data points** *refers to text or a label in the ESCI page.*

3.5 Use the spinner near red 3 to select your sample size, perhaps $N = 20$ or whatever you choose. Take a deep breath then click the **Take sample** button. You should see something like Figure 3.1. The scatter of points at the bottom is the 20 data points of our sample. That's our simulated equivalent of finding a random sample of N students and testing them on the HEAT.

3.6 Take more samples. The scatters of data points for successive samples vary greatly. As we'll discuss in later chapters, randomness is intriguing and weird. It's important to develop good intuitions about randomness, although this is a challenge because there's evidence that people often underestimate the extent of random variability.

3.7 Observe your samples of data points carefully. Would you agree that the sampled data points in the long run are about equally often below and above μ? That they tend to cluster fairly close to μ, but values farther from μ are quite common? Just occasionally you get an extreme point? Those features of sampled values follow directly, of course, from the shape of the population and our sampling, which we assume gives every data point in the population an equal chance of being chosen.

Sampling: The Mean Heap and Dance of the Means

The next important idea is the *sampling distribution of the sample mean.* Imagine taking lots of samples from a population—we'll do that with ESCI in a moment. The means of those successive samples vary, but they tend to cluster around the population mean μ. Many such sample means form a distribution, called the sampling distribution of the sample mean. If we could take an infinite number of samples, their means would form a normal distribution, thus demonstrating that the sampling distribution of the sample mean is normal. It's an excellent question why this sampling distribution is normal in shape. The general answer to that question reveals a little magic.

The *sampling distribution of the sample mean* is the distribution created by the means of many samples.

The previous paragraph is an example of explanation that will shortly be followed by ESCI activities. If the text there is unclear, read on, and return after exploring the ideas with ESCI. Or peruse the text, and the ESCI exercises that follow, in parallel.

Some Statistical Magic, the Central Limit Theorem

You might wonder why statisticians choose the normal distribution as a statistical model. The answer is the *central limit theorem,* which is the central result in all theoretical statistics. If you make a variable (let's call it X) by adding up lots of other variables, all independent, then X has, at least approximately, a normal distribution. The amazing thing is that it has a normal distribution pretty much whatever the distributions are of the other variables you add to get X. All those variables can even have distributions of different shapes but, provided they are all independent, their sum X is approximately normally distributed. If more variables are added, X is closer to normal. The normal distribution appears out of thin air, and in this way represents some fundamental aspect of the universe.

Shortly we'll use ESCI to illustrate that the sampling distribution of the sample mean is normally distributed. ESCI currently offers only normal populations, but a future version might offer populations with distributions other than normal—maybe skewed, or with more than one hump, or different in other ways. The amazing thing is that even populations with weird distribution shapes will give a sampling distribution of means that's approximately normal, and closer to normal for samples with larger N. We're talking about two distributions here—the population and the sampling distribution of means. The central limit theorem states that the latter is approximately normal in shape, almost regardless of the shape of the population distribution.

Think of the sample mean as the sum of lots of tiny, independent contributions—the data points in the sample. The central limit theorem states that such sums have approximately a normal distribution. This way of thinking about the theorem gives a link with nature. Suppose you measure some natural quantity in the world, such as the length of adult ants or the time it takes for penguin eggs to hatch. Fairly often, although not always, a large set of such measurements is approximately normally distributed. If ant length or hatching time is determined by the addition of numerous separate influences—perhaps genetic, environmental, nutritional, or random—then the central limit theorem says the result will be approximately normal. No doubt mere addition of independent influences is much too simplistic a biological model, but the idea probably does explain why the normal distribution often appears in nature, at least approximately. The central limit theorem and the normal distribution do seem to express some basic aspects of how the natural world functions.

The Standard Error

The SD of the sampling distribution of the sample mean is called the *standard error* (SE). That may be confusing, so it may be worth making it a chant. Dismay your friends at parties by intoning: "*The standard error is the standard deviation of the sampling distribution of the sample mean.*" You can easily explain by pointing to the mean heap—which we'll discover in a moment. We'll use ESCI to picture the SE, and to illustrate the formula:

A chant: "The *standard error* is the standard deviation of the sampling distribution of the sample mean."

$$SE = \sigma/\sqrt{N} \tag{3.1}$$

which is one of the few formulas you need to explore and remember.

We use ESCI to run simulations, which can be revealing. However, a simulation is not real life. It is vital to keep in mind two major ways that ESCI simulations differ from the usual research situation:

1. A simulation requires that we assume some particular population distribution. You choose a normal distribution, and values of μ and σ, which are shown on the screen. In our role as real-life researchers, however, we never know μ or σ—we are running the experiment to estimate them.
2. We usually take many simulated samples, whereas in real life we almost always can run an experiment only once.

Distinguish carefully between playing around on the computer with simulations of many experiments, and running and analyzing a single experiment in real life.

Exercises

3.8 We'll now work toward generating pictures like those shown in Figure 3.2. Click **Clear** (button near red 1), and click near red 3 to **Display means**. Take a sample. The sample mean is displayed as a green dot just below the scatter of data points.

3.9 Click near red 4 to **Show values**. Values are shown on screen for M and s, the sample statistics for the latest sample you've taken. We can compare these values with the values we've chosen for their population counterparts μ and σ.

"Near red 4" can refer to anywhere in the colored area that has red 4 at its top left corner.

3.10 Click **Take sample** a few times. The means drop down the screen, as in Figure 3.2. Watch the values bounce around, and compare them with the μ value you set. Each click is equivalent to running an experiment, meaning you take a new sample of size N, obtain the HEAT scores for those N students, and then calculate M and s to use as estimates of μ and σ, the unknown parameters we're studying.

3.11 Click **Run-Stop** and watch the sample means dancing down the screen. It's the *dance of the means*, as in Figure 3.2, which illustrates the extent of variation or bouncing around of the mean from sample to sample. (If the dance is a bit slow, try clicking near red 2 to hide the population.) Imagine (or play on your computer) your choice of backing music for the dance.

The *dance of the means* is my name for a sequence of sample means falling down the screen.

3.12 Click **Run-Stop** again to stop the dance, then **Clear**. Now think about two predictions: First, if you change N, what will happen to the dance? For example, will larger N give a more drunken dance—the means tending to vary side-to-side more—or a more sober dance? What about smaller N? Make your predictions—write them down. The two halves of Figure 3.2 illustrate

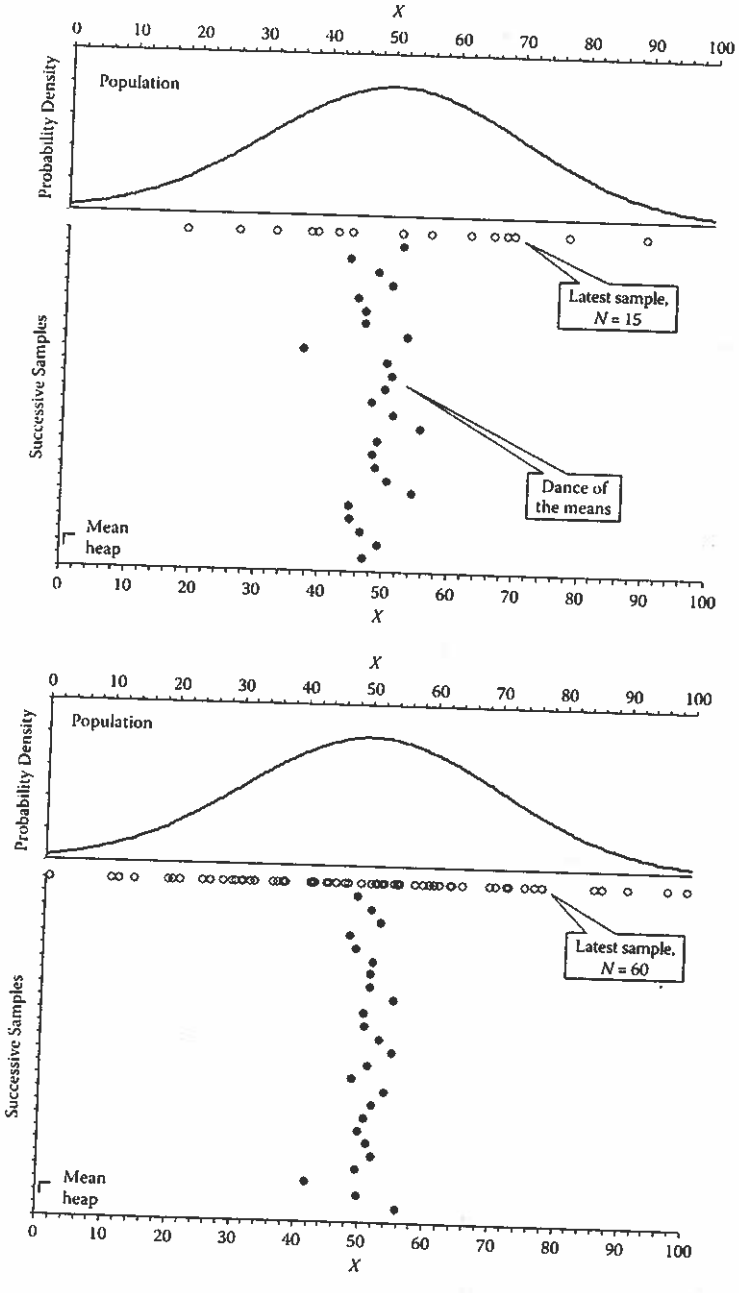


FIGURE 3.2
Dance of the means—the dots dropping down the screen. Upper half: $N = 15$. Lower half: $N = 60$. In each case the population distribution is displayed at the top, and the latest sample appears as the scatter of N data points in a horizontal line just below.

the dance for different values of *N*. Are your predictions consistent with what that figure shows?

3.13 Second, what would happen if you increase or decrease σ? Any change to the drunkenness of the dance? Which way would it change? Lock in your predictions.

3.14 Experiment to test your predictions. Which change—a different *N* or a different σ—tends to make more difference?

3.15 Click **Mean heap** near red 5 to see all the means collapse down into a pile of green dots, as Figure 3.3 illustrates. This is the sampling distribution of the mean, and I call it the *mean heap.* Run the simulation to build up a good-sized heap. Do this for various values of *N*, and keep track of how wide the heap appears: Record for each *N* your eyeball estimate of the SD of the mean heap. (It may help to recall the rule of thumb that about 95% of the values in a normal distribution lie within 2SD on either side of the mean. If that's unfamiliar, explore Appendix B.) Figure 3.3 shows the mean heap for two values of *N*. Should we prefer a narrow or a wide mean heap, bearing in mind that we are trying to estimate μ? Translate your conclusion into advice for a researcher who is considering what size sample to take.

The *mean heap* is my name for the samp distribution of the sample mean. It's a of green dots that represent sample mea

In Figure 3.3, the mean of the latest sample, which has just been added to the mean heap, is highlighted as a large black dot. In ESCI it appears as a dark green dot, the same size as the dots for the other means. Small features sometimes appear to be a little different in the figures than in ESCI, to clarify what the figures show.

3.16 Click **Display sampling distribution curve** near red 6. The normal distribution displayed on the mean heap, as in the lower panel of Figure 3.4, is the *theoretical sampling distribution of the sample mean.* We can compare that with the mean heap, which is the *empirical sampling distribution of the sample mean*—the heap of just the means we've taken so far. The curve is the distribution theoretically predicted from knowing μ, σ, and *N*. (In ESCI, the curve is scaled vertically so it fits to the mean heap. Take more samples, and both the mean heap and sampling distribution curve grow higher—but not wider; the SD of the sampling distribution remains the same.)

Take an infinite number of samples and distribution of their means is the *theoret sampling distribution of the sample m* (The mean heap is my name for the *empi sampling distribution of the sample mea*

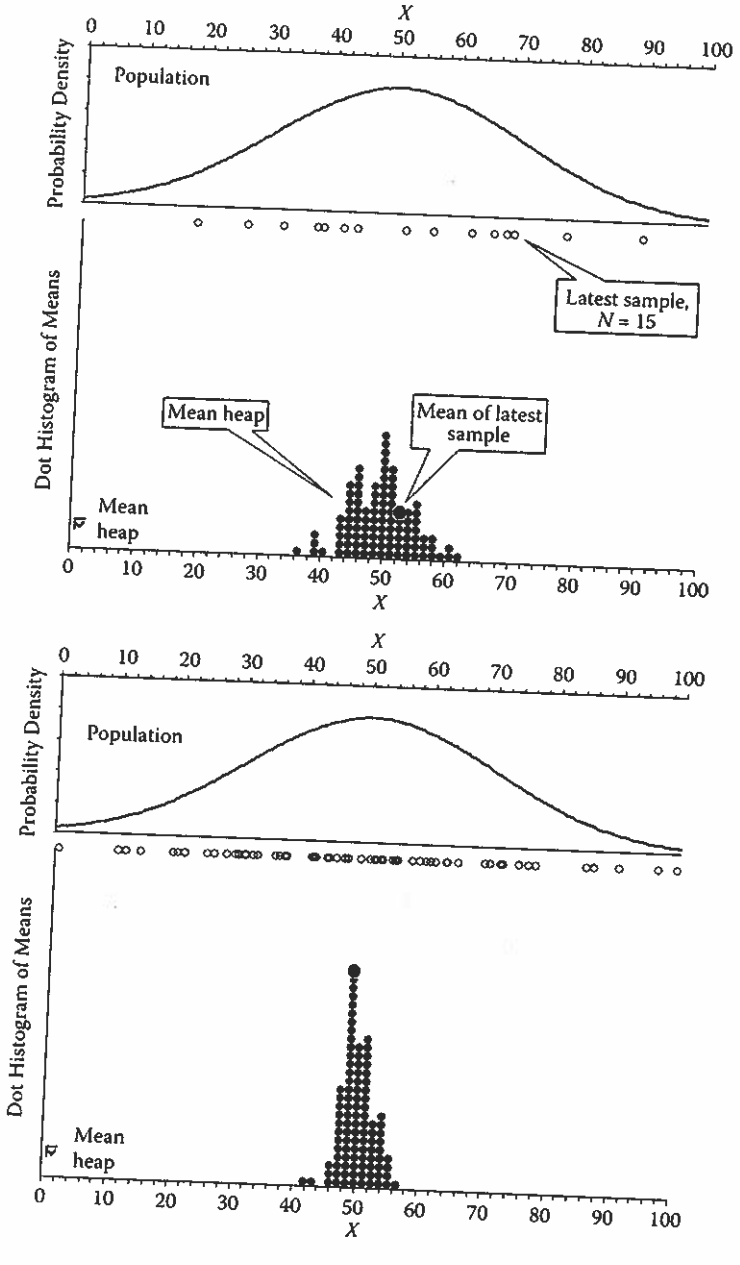


FIGURE 3.3
The mean heap, in each case after taking 100 samples. Upper half: $N = 15$, and the eyeball estimate of the SD of the mean heap may be about 5. Lower half: $N = 60$, and the SD of the mean heap looks to be about 3. The mean of the latest sample is displayed as a highlighted dot when it is added to the heap.

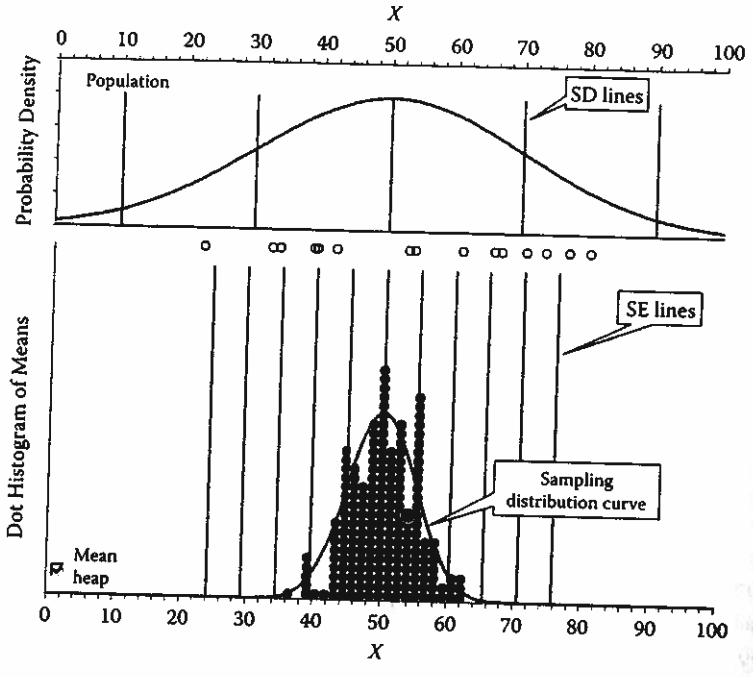


FIGURE 3.4

The upper panel displays the population distribution, with lines marking SD units, showing $\sigma = 20$. Below is the mean heap. The superimposed curve is the sampling distribution of the mean, with lines marking SE units. In this example, $N = 15$, and 200 samples have been taken. The SE = $\sigma/\sqrt{N} = 20/\sqrt{15} = 5.16$.

3.17 The SD of the sampling distribution of the sample mean is given the name *standard error* (SE), so the SE is just a particular SD. Think of the SE as summarizing the breadth or spread of the mean heap—or its curve.

3.18 Click **Display SE lines** near red 6 and see vertical lines marking SE units across the sampling distribution curve. These vertical lines are displayed in the lower panel of Figure 3.4.

3.19 Near red 6 find **Curve SE** and note its value. The popout comment explains that it's the SE of the sampling distribution curve. Does it change if you take further samples? Why?

3.20 Click **Display SD lines** near red 2 (if necessary, click **Display population** first). SD lines for the population curve are displayed, as in the upper panel of Figure 3.4. Compare these with the SE lines for the sampling distribution curve in the lower panel. In each case you can regard the lines as marking $z = 0$ (the mean), and $z = -2$, -1, and $+1$, $+2$, etc., for the respective

normal distributions. (Click **Display means** near red 3 to hide the means, if you want to see the sampling distribution curve more clearly.)

3.21 The sampling distribution is normally distributed—as the curve on the mean heap illustrates—with mean μ and SD of $\sigma/\sqrt{N}$. The vital formula to remember is SE = $\sigma/\sqrt{N}$. Maybe write this formula down, for safekeeping. The mean heap, and its curve, is centered symmetrically under the population, and its SD is smaller than that of the population—by a factor of $\sqrt{N}$.

3.22 If N is made four times bigger, $\sqrt{N}$ becomes twice as large, so the SE should be halved. Compare the lower halves of Figures 3.2 and 3.3, for which $N = 60$, with the upper halves, for which $N = 15$. Does the lower dance seem about half as varied, half as wide as the upper? The lower mean heap about half as wide as the upper? Unfortunately, to halve the amount of variation we need to take a sample four times as big. That's bad news for researchers trying to make precise estimates because, as we'll see, the SE determines precision. A broad mean heap signals a large SE and imprecise estimates.

3.23 This might be a good spot to cross-check with any other statistics textbook you are using. See if you can use ESCI to illustrate the way your other textbook explains sampling and sampling distributions.

3.24 Use the values of σ and N that you set, and which are shown near red 2 and red 3, to calculate SE. Check that the value shown at **Curve SE** is correct.

3.25 Suppose HEAT scores have mean = 50 and SD = 20 in your country. For samples of $N = 30$, what is the SE? (Use the formula to calculate it, then use ESCI to check.) Describe the sampling distribution of the mean.

3.26 That's a typical textbook problem. Invent and solve a few more. Do a few from your other textbook. Maybe invent some, and swap with a fellow learner.

3.27 Recall our chant: "The standard error is the standard deviation of the sampling distribution of the sample mean." If someone asks, "What's a standard error?" you can bring to mind the mean heap as a pile of green dots, then explain about its SD.

3.28 Make up some exercises for discovery learning of the SE = $\sigma/\sqrt{N}$ relation. You could suggest first making predictions, or guesstimates, of the SE of the mean heap (and the sampling distribution curve) for a few widely separated values of N that you nominate. Then, for each of those N values, take at least 50 samples

and eyeball the SD of the mean heap—which as you know is the SE. See Figure 3.3 and its caption. Compare those eyeballed estimates with the ESCI values near red 6 for the **Mean heap SE**, which is the SE of the displayed mean heap. What does a graph of those SE values against N look like? How accurate were the original predictions? Find someone who doesn't know about SE to try out your exercises.

Errors of Estimation, and the Margin of Error

We take a sample and calculate M because we want an estimate of μ. How good an estimate is it? The *estimation error* is $(M - \mu)$, and is different for every sample. The center of the mean heap is at μ, and the sample means, shown by the green dots, cluster around μ but generally fall a little to the right or left of μ. The distance away they fall is $(M - \mu)$, the estimation error. We can think of the mean heap, and the sampling distribution of M, as the distribution of estimation errors. Most green dots fall fairly close to μ, so have small estimation errors; many fall a moderate distance away; and just a few fall in the tails of the sampling distribution, which signals large estimation errors.

Estimation error is $(M - \mu)$, the distance between our point estimate based on the sample and the population parameter we are estimating.

We define the *margin of error* as the largest likely estimation error. The abbreviation is MOE, which you can read out as M-O-E, although I prefer to say it as "MOW-ee." We usually choose "likely" to mean 95%, so there's a 95% chance that the estimation error is less than the MOE, and only a 5% chance that we have been unlucky and our sample mean M falls in one of the tails of the sampling distribution. You probably know the rule of thumb for any normal distribution: About 95% of the values fall within 2SD on either side of the mean. Therefore, 95% of sample means will fall within about 2SE of the mean of the sampling distribution. (Remember that SE is the SD of that distribution. I sometimes suspect that those terms were selected to be as confusing as possible.) We can therefore state that MOE = 2SE, approximately, and that's the value to remember for eyeballing purposes. More accurately, MOE = 1.96 × SE because 1.96 is the critical value $z_{.95}$ from a normal distribution. (Appendix B does some relevant explaining.)

The *margin of error* (MOE) is the largest likely estimation error. If "likely" is taken to mean 95%, MOE is approximately 2SE.

The 95% of the M green dots that fall within MOE (i.e., about 2SE) on either side of μ have estimation error less than MOE, and only the 5% that fall farther than this from μ, within one or the other tail, have

estimation error greater than MOE. Warm up the trumpets: The CIs are about to arrive.

Exercises

3.29 With the mean heap, sampling distribution curve, and SE lines displayed, click near red 6 to **Display ±MOE around μ**. Your screen should resemble Figure 3.5. On the bottom axis is a green stripe that indicates a distance of one MOE on either side of μ. At the ends of the stripe, heavier green vertical lines mark a distance of MOE on either side of μ. How many SE units away from μ are the heavier green vertical MOE lines? Is that what you expected?

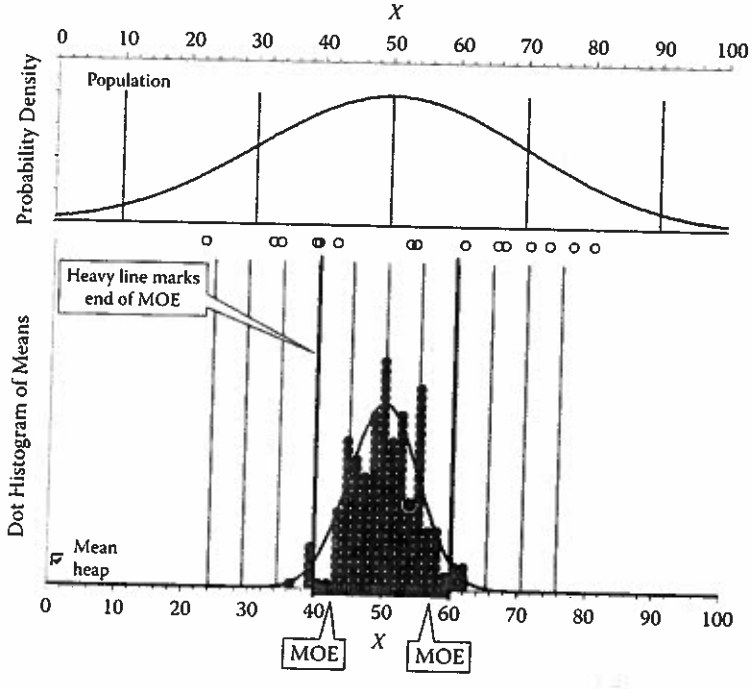


FIGURE 3.5
Same as Figure 3.4, but with MOE marked. The stripe at the bottom, which is green on the screen, extends MOE on either side of the mean μ = 50, where MOE = 1.96 × 5.16 = 10.12. The ends of the stripe are marked with heavier vertical lines. We expect about 95% of means to fall between those lines and, correspondingly, we expect 95% of the area under the sampling distribution curve to fall between those vertical lines.

3.30 What percentage of green dots will fall between those MOE lines? What percentage beyond those lines to the left? Beyond them to the right?

3.31 Click near red 5 to return from the mean heap to the dance of the means, run the simulation, and watch how often a mean falls outside the MOE lines, to the left or right.

3.32 Suppose HEAT scores have $\mu = 50$ and $\sigma = 20$. For $N = 36$, calculate the MOE (i) approximately, and (ii) exactly. Use your answer to (ii) to find an interval that in the long run should include 95% of sample means.

3.33 Set up that situation in ESCI, make sure **Display ±MOE around μ** is clicked on, and note the MOE value shown near red 6. Check that it's the same as you calculated.

3.34 How would you expect MOE to change for different N? For different σ? Test out your predictions. In each case, note about how many green dots fall outside the MOE lines.

3.35 Consider our initial question about the mean HEAT scores in your country. State your aim in an estimation-thinking "how much" way.

CIs at Last: Sound the Trumpets!

We've talked about MOE as describing how sample means clump around μ. Informally, MOE tells us about the "width" of the mean heap, or of the sampling distribution: The green stripe at the bottom of the mean heap, as in Figure 3.5, is 2MOE long and includes 95% of means. In 95% of cases the estimation error is less than MOE, or in other words $|M - \mu| < \text{MOE}$. (The vertical bars mean absolute value, so $|M - \mu|$ equals whichever of $(M - \mu)$ and $(\mu - M)$ is greater than zero.)

Figures 3.1 to 3.5 show simulations, in which we assume μ and σ are known. Now consider Figure 3.6, which shows all we know as typical researchers: our single sample of $N = 15$ data points and their mean. All this ESCI work with simulations is intended to build intuitions about what lies behind such a set of data. Whenever you see a data set, first bring to mind the population and recognize that you don't know its μ or σ. In practice you usually also don't even know whether or not the population is normally distributed, although here we're assuming it is. Next, visualize the dance of the means and the mean heap. We have a single green dot, but it's randomly chosen from the infinite dance. The drunkenness of

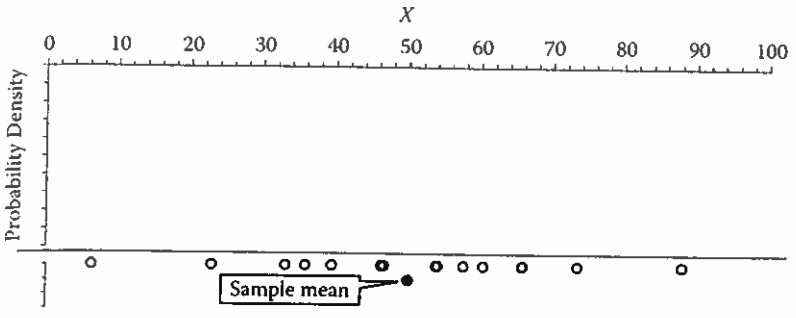


FIGURE 3.6
All a researcher knows: a single sample of $N = 15$ data points and their mean.

the dance—meaning the amount the means bounce around from side to side—or the width of the mean heap, would tell us how far our M might be from the μ we want to estimate. As well as thinking of their own data, researchers should always think about what other values they easily could have obtained, if they'd happened to take a different sample. We could have obtained any of the means in the dance.

We'll use our M to estimate μ, and we want to know how good an estimate it is. Here's a wonderful fact: We can use MOE to provide information about that precision. We rely on this obvious relation: If M is likely to be close to μ—as the last page or two has illustrated—then μ is likely to be close to M. As simple as that. The simulation shows us that, for most samples, M falls pretty close to μ, in fact within MOE of μ. Now we have only a single M and don't know μ. But, unless we've been unlucky, our M has fallen within MOE of μ, and so, if we mark out an interval extending MOE on either side of our M, most likely we've included μ. Indeed, and that interval is the *confidence interval* (CI)!

We define the interval $[M - \text{MOE}, M + \text{MOE}]$ as the CI. In 95% of cases, that interval will include the unknown population mean μ. That's the interval we want, and so we can now celebrate with the trumpets. Recall that MOE = 1.96 × SE = $1.96 \times \sigma/\sqrt{N}$. Therefore,

confidence interval on the sample mean
s the interval $[M - \text{MOE}, M + \text{MOE}]$,
:h extends MOE on either side of M.

$$\text{the 95\% CI is } [M - 1.96 \times \sigma/\sqrt{N}, M + 1.96 \times \sigma/\sqrt{N}] \qquad (3.2)$$

For eyeballing purposes, use 2 in place of 1.96.

We can label the 95 as the *level of confidence*, C, because it specifies how confident we can be that a CI includes μ. It's also referred to as the *confidence level*. We usually choose $C = 95$, but other values are possible and you can use ESCI to experiment with them.

The *level of confidence*, or *confidence level*, is the 95 in "95% CI." It specifies how confident we can be that our CI includes the population parameter μ.

You might have noticed a problem: MOE is calculated from σ but, you ask, how can we do that when we don't know σ? You're correct, but as a first step we'll assume σ is known and use it to calculate MOE and the CI. As a second and more realistic step, we'll use our sample s as an estimate of σ in our calculation of MOE for the CI. Assuming σ is known, MOE is calculated using $z_{95} = 1.96$. Dropping that assumption and using s to estimate σ, we need instead to use a critical value of t. As you probably know, using t requires us to choose an appropriate value for the *degrees of freedom* (*df*). For our situation, with a single sample, $df = N - 1$, and the critical value we need is $t_{95}(N - 1)$. Use the **Normal z t** page of **ESCI chapters 1–4** and the notes in Appendix B to find any critical values of z or t that you need. Then,

$$\text{the 95\% CI is } [M - t_{95}(N-1) \times s/\sqrt{N},\ M + t_{95}(N-1) \times s/\sqrt{N}\,] \quad (3.3)$$

Exercises

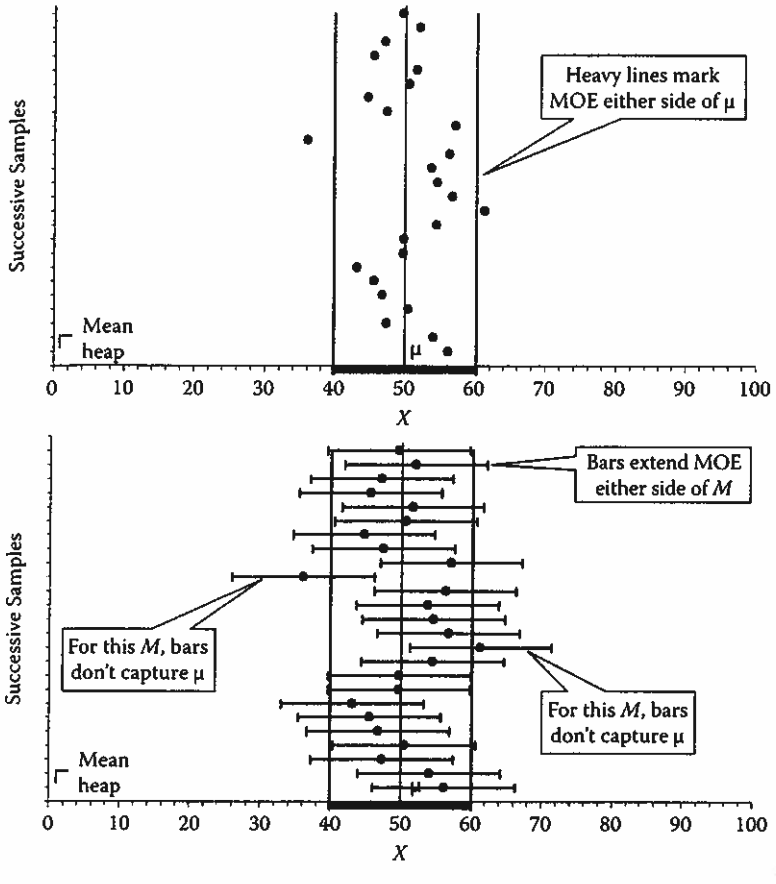
3.36 Display the dance of the means, click near red 8 to mark μ with a black vertical line, and click **Display ±MOE around μ**. Compare with Figure 3.7, upper half. Do you have any means beyond MOE? What percentage would you expect in the long run?

3.37 The green stripe at the bottom has length 2MOE. We are going to take a line of that length and place it over each mean to mark an interval extending MOE on either side of the mean. (At this point, make sure that **Assume σ known** near red 7 is clicked on, but **Mean heap** is not clicked.) Near red 7 click **Display CIs**, and there they are. Run the simulation and enjoy the *dance of the CIs*. Music? Compare with Figure 3.7, lower half.

The *dance of the confidence intervals* is a sequence of CIs bouncing around for successive samples, as in Figure 3.7, lower half, and Figure 3.8.

3.38 A CI *includes* μ, or *captures* μ, every time, unless the mean falls outside the MOE lines. Run the simulation and watch. What percentage of CIs will in the long run miss μ? What percentage will miss to the left? To the right?

3.39 All our CIs are the same length because we are using the same MOE value for each. That's calculated from σ, which for the moment we're assuming is known.



URE 3.7

ıce of the means, $N = 15$ for each sample. A vertical line marks $\mu = 50$. In each half of the ıre a stripe at the bottom extends MOE on either side of μ, and the ends of the stripe are rked by vertical lines. Two means happen to fall outside those MOE lines. In the lower f, error bars of length MOE on either side of each mean are displayed: These are the 95% . Only for the two means falling outside the MOE lines does the CI fail to include μ.

3.40 Unclick **Display ±MOE around μ** to hide the MOE lines; then near red 9 click **Show capture of μ**, as in Figure 3.8, upper half. If a CI doesn't capture μ, ESCI displays it in red. Do you have any red CIs? Explain.

3.41 Click off **Assume σ known** near red 7. What happens? Compare with Figure 3.8, lower half. Click on and off a few times and watch carefully. If we drop the assumption that σ is known we are being much more realistic. MOE is now calculated using s as our estimate of σ.

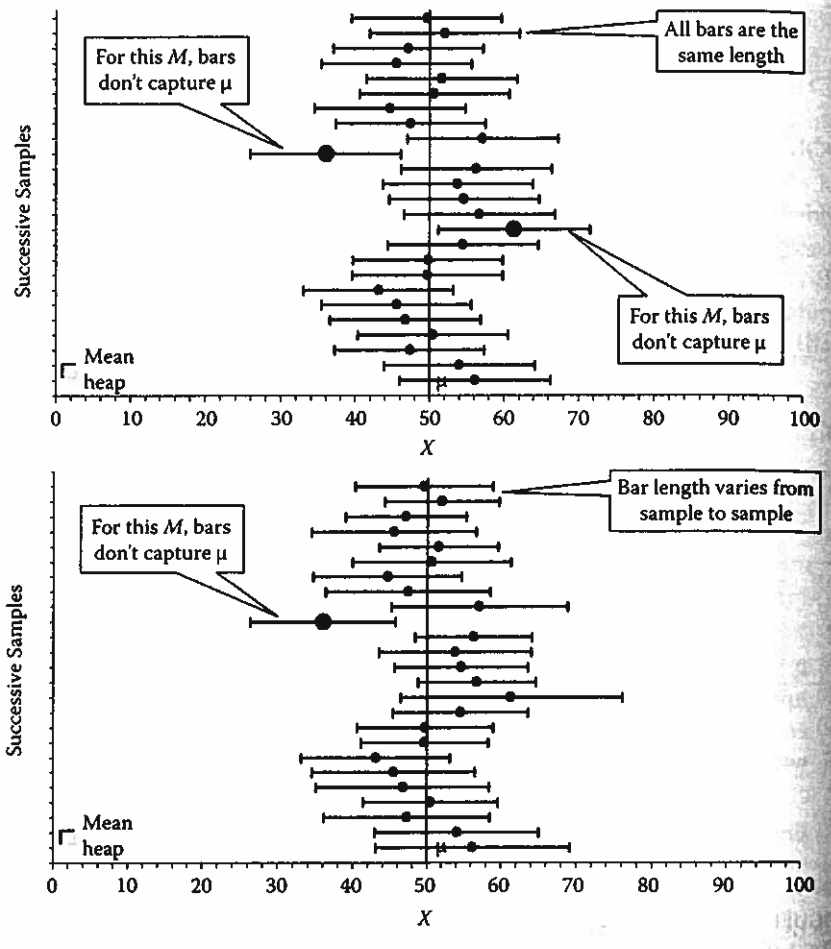


FIGURE 3.8
Dance of the CIs, $N = 15$ for each sample. CIs that miss μ are marked here with a larger black dot for the mean; in ESCI they are red. Upper half: Assuming σ is known. Lower half: That assumption is dropped and each CI is calculated using s for that sample, so the CIs vary in length. Whether a CI captures μ or not may change when the assumption about σ changes, as here for the sample 11th from the bottom.

3.42 Every sample has its own s, and so the CIs vary in length from sample to sample. What would happen for $N = 10$, or even smaller? Would s vary more or less, from sample to sample? Would s typically be a better or worse estimate of σ? Would you expect CI length to vary more, or less, from sample to sample? What about $N = 100$?

3.43 Experiment to test your predictions. Explain.

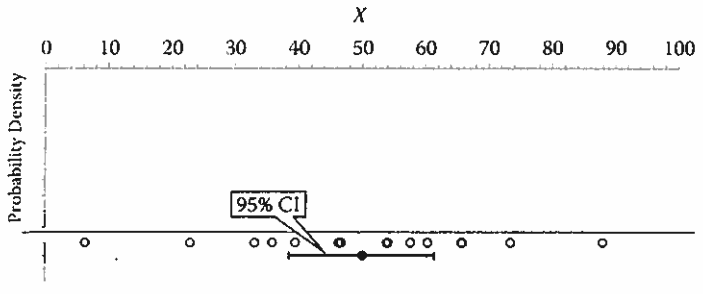


FIGURE 3.9
All that the researcher knows: the data points of a single sample with $N = 15$, as shown in Figure 3.6, but now the 95% CI has been calculated, using s.

3.44 Assuming σ is known, the CI was based on σ and the critical value of z. Without knowing σ, we use instead s and the critical value of t. So the 95% CI is $[M - t_{.95}(N-1) \times s/\sqrt{N}, M + t_{.95}(N-1) \times s/\sqrt{N}]$. Suppose you choose $N = 20$, take a sample, and calculate from your data that $M = 44.2$ and $s = 17.5$. Calculate the 95% CI. Use the **Normal z t** page of **ESCI chapters 1–4** to find the critical value of t you need; Appendix B has suggestions to help. While you are using that ESCI page, have a play with the shapes of the t distribution for various values of df, and compare with the normal distribution.

3.45 Calculating that CI is a typical textbook problem. Invent a few more, as varied as you can, and swap with a fellow learner. Work out the answers to each other's problems. (That's a good thing to do with a friend?) Do some similar exercises from another statistics textbook you know.

3.46 Figure 3.9 is the same as Figure 3.6, but now we have calculated the 95% CI using s from our data points. How should we think about that CI?

What does the 95% mean? In general, how should we interpret a CI? We'll discuss that in future chapters, but the figures around here give the basic answer. Our CI is one of an infinite sequence of possible CIs generated from the infinite dance of the means—from the infinite collection of samples, any of which we might have obtained in our experiment. In the long run, 95% of those CIs will capture μ, and 5% will miss. CIs that miss are shown in Figure 3.8 with a larger black dot, and in ESCI are shown in red.

It's a basic CI slogan: "It might be red!" We can be 95% confident that our CI captures μ, but it might be red. In your lifetime of calculating and reading and considering numerous

For any CI bear in mind, "It might be red!" It might be one of the intervals that don't capture μ, although in real life we'll never know.

95% CIs, around 95% will include the population parameters they estimate, and 5% will be red. It's a great convenience that ESCI can display in red the CIs that miss μ. Alas, in real life CIs don't come in color: You can never be sure whether any particular CI should be red or not.

3.47 Return to ESCI **CIjumping** and set up the dance of the CIs, showing capture, as in Figure 3.8. What do you expect if we change C, the level of confidence? Would 99% CIs be narrower or wider than 95% CIs? You are aiming for higher confidence of capturing μ, so would you need a narrower or a wider net? What about 90% or 80% CIs? Lock in your predictions.

3.48 Near red 7 is the spinner to set C. Read the popout. Change C and test your predictions. Does it make sense that CIs sometimes change color as you change C? (*Note*: The spinner will give you values up to 99, but you can type in values up to 99.9. Type in a value, then press Enter on your keyboard.)

3.49 Play around with C. Think back to Figure 3.5, the mean heap, and MOEs and the percentage of the mean heap they include. Any surprises as you vary C over a wide range?

3.50 Set C = 95, the value we almost always use. Click **Assume σ known** on, so CIs are all the same length. Make sure capture of μ is indicated by deep green or red. Run the simulation, enjoy the dance of the CIs to your favorite backing music, and watch **Percent capturing** near red 9. What happens near the start of a run? What happens after a minute or two? After 10 minutes? After an hour or more?

3.51 Do it all again without assuming σ known, so the CIs vary in length.

3.52 Do it all a few more times, with various values of N including some very small values, and N = 100, the maximum this simulation allows.

3.53 Do it all yet again for various values of C.

Early in a run, after taking a small number of samples, the percentage capturing may differ a bit from C. Do you find it impressive that, after a minute or two, and certainly after 10 minutes or more, the percentage capturing is close or very close to C? Even more impressively, that's true for any N, any C, and whether or not you assume σ is known. Yes, the formulas for CIs predict extremely well how random sampling behaves. (And the random number generator I am using in ESCI is very good.)

3.54 Study Figure 3.9 again. That's all a researcher knows. That's all that's available for the Results section in a journal article. Whenever you see such a figure, or a Results section, you should bring to mind two underlying things to illuminate how you think about the results. What are they? *Hint*: Look back to Figure 3.6 and the discussion of that.

Reporting CIs

This is a good moment to reflect on the journey so far. I've argued that NHST is deeply flawed, is often misunderstood, and can mislead. If used well, estimation not only avoids the problems of NHST, but prompts researchers to ask "how much?" questions, and these questions are likely to give more informative answers than NHST's dichotomous questions. Estimation is based on point ES estimates, and interval estimates—which are the CIs we've just been discussing. So we've now encountered ESs and CIs, which are the basic building blocks for the new statistics. The main business of the rest of this book is to consider additional ESs, discuss six ways to interpret CIs, and introduce meta-analysis—which itself is based on ESs and CIs.

It's worth celebrating CIs and what they have to offer. (Wine, coffee, or another play with **CIjumping**?) The simple message of this section is that we should always, where possible, report CIs for any ES estimates. Report them in the text, in tables, or in figures, as is most helpful for your readers. I'll now outline what the *Publication Manual* (APA, 2010) says about reporting CIs.

Recommendations of the APA *Publication Manual*

Your discipline may not use the *Publication Manual*, but journals in a very wide range of NHST disciplines refer to it, so its advice is influential. In any case, it says sensible things about CIs. It includes a strong statement about CIs:

> The inclusion of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results.... Confidence intervals combine information on location and precision ... they are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended. (APA, 2010, p. 34)

The *Manual* then goes on to make the key recommendation I have already mentioned that researchers should "wherever possible, base discussion and interpretation of results on point and interval estimates" (p. 34).

The *Manual* specifies that a CI reported in text should be shown in the square bracket format I have been using in this book: $z = 0.65$, 95% CI [0.35, 0.95]. Here's another example: $M = 30.5$ cm, 95% CI [18.0, 43.0], which shows that the units of measurement (cm) should not be repeated in the square brackets. You might think the simple matter of how to report a CI in text would have been decided years ago, and some nice, clear format would be everyone's choice. Surprisingly, however, not even medicine has settled on a format that has become widely used, even though it has routinely reported CIs since the 1980s. I hope the *Manual's* [..., ...] format will quickly become familiar as signaling a CI. The *Manual* uses the same format for other levels of confidence, so you could, for example, report a mean response time as $M = 625$ ms, 99% CI [564, 686]. My recommendation, though, is to use 95% CIs unless there are good reasons for choosing some other level of confidence. It's challenging enough to build up good intuitions about the standard 95% level of confidence without trying to cope with CIs having a variety of levels. There's more about that in Chapter 4.

To report a CI, use this format: "The response time was $M = 567$ ms, 95% CI 622]." For subsequent intervals, you omit the "95% CI" if the meaning is cl

The *Manual* says the "95% CI" (or other level of confidence) should appear before the square brackets for the first CI reported in any paragraph, but can be omitted for any further CIs reported in the same paragraph. So a later report in the same paragraph might be $z = 1.12$, [0.66, 1.58]. However, I hope the [..., ...] format will become so familiar for 95% CIs that the "95% CI" need only be stated once at the start of an article.

Chapter 5 in the *Manual* includes many examples of tables, and it's excellent that four of those examples include CIs. The wise advice is, "When a table includes point estimates, for example, means, correlations, or regression slopes, it should also, where possible, include confidence intervals" (APA, 2010, p. 138). To show CIs in a table you can either use the [..., ...] format, or use separate columns for the values of the interval endpoints: the lower limit (LL) and upper limit (UL) of each interval.

Tables 3.1 and 3.2 illustrate those two formats recommended by the *Manual*, using a small proportion of the data reported by Strandberg-Larsen, Grønbæk, Andersen, Andersen, and Olsen (2009). The tables show point and interval estimates for the relative risk of postneonatal mortality, defined as an infant dying between 28 days and one year after birth, for various levels of drinking by the mother. The risk for mothers reporting no drinking is used as the comparison, so the first row of data shows

TABLE 3.1

An Example Table Reporting the Association of Alcohol Consumption During Pregnancy With Infant Postneonatal Mortality

Alcohol Consumption (Average Drinks/Week)	Relative Risk of Postneonatal Mortality (Adjusted)	95% CI
0	1.00	—
0.5–1.5	0.82	[0.48, 1.39]
2–3.5	0.68	[0.27, 1.71]
4 or more	2.91	[1.22, 6.95]

Source: Data from K. Strandberg-Larsen, M. Grønbæk, A.-M. N. Andersen, P. K. Andersen, & J. Olsen (2009). Alcohol drinking pattern during pregnancy and risk of infant mortality. *Epidemiology, 20,* 884–891.

TABLE 3.2

A Second Example Table Reporting the Association of Alcohol Consumption During Pregnancy With Infant Postneonatal Mortality

Alcohol Consumption (Average Drinks/Week)	Relative Risk of Postneonatal Mortality (Adjusted)	95% CI	
		LL	*UL*
0	1.00	—	—
0.5–1.5	0.82	0.48	1.39
2–3.5	0.68	0.27	1.71
4 or more	2.91	1.22	6.95

Source: Data from K. Strandberg-Larsen, M. Grønbæk, A.-M. N. Andersen, P. K. Andersen, & J. Olsen (2009). Alcohol drinking pattern during pregnancy and risk of infant mortality. *Epidemiology, 20,* 884–891.

relative risk of 1.00. The values of relative risk are adjusted to allow for differences in many other characteristics, including maternal age, socioeconomic status, and smoking status. The tables suggest that levels of drinking up to an average of 3.5 drinks per week have little impact on postneonatal mortality, but a mean consumption of four or more drinks per week increases risk. The point estimate is an increase by a factor of about 3 but, despite the study analyzing data for about 80,000 births, the CI is wide—from about a 20% increase in risk to an increase by a factor of about 7. Having the CIs certainly gives a fuller picture of the results than either just the point estimates, or those estimates plus the information that only for the bottom row does the increase reach statistical significance.

Interpreting CIs

In this and the next two chapters I'll describe six ways to think about and interpret CIs. The first is based on the definition of a CI. The following five I'll describe in an order that's convenient for presentation, but it is not an order of priority or preference. It's helpful to have many possibilities in mind for interpreting CIs, and to use whichever one or ones are most illuminating in a particular situation. After I've discussed them all, Table 5.1 provides a summary.

It may seem surprising, but the experts are not fully agreed on how best to interpret CIs. Of the six approaches I'll describe, only the first, which is based on the definition of the level of confidence, is fully endorsed by everyone. The others each attract quibbles or criticism from one or another expert. I'll explain something about some of the issues, but mainly I'll take a pragmatic approach and discuss approaches to interpretation that seem valuable to me. All six interpretations I describe are, in my view, reasonable as well as often useful in practice.

CI Interpretation 1: One From the Dance

Interpretation 1 of a CI. Our CI is one the *dance*—an infinite sequence of re of the experiment. Most likely it capture parameter we're estimating, but, "It be red!"

As I mentioned, it's always correct to think of the CI calculated from our sample data as one from the potentially infinite sequence of intervals that we'd obtain if the experiment were repeated indefinitely. Each interval is just one randomly chosen from the dance of the CIs. As you inspect your interval, have in your mind's eye the dance of similar intervals. You realize that if your interval is narrow, most likely the dance is quite sober, but the wider your interval, the more varied (and wider) the dance is likely to be. Most likely your interval captures the parameter you wish to estimate, but, by chance, you may have an interval that doesn't include the parameter and that ESCI would show in red. Never forget, "It might be red!" Example 3.1 refers to the first interpretation of a CI.

CI Interpretation 2: Interpret Our Interval

It's tempting to say that the probability is .95 that μ lies in our 95% CI. Some scholars permit such statements, while others regard them as wrong, misleading, and wicked. The trouble is that mention of probability suggests μ is a variable, rather than having a fixed value that we don't know. Our interval either does or does not include μ, and so in a sense the probability is either 1 or 0. I believe it's best to avoid the term "probability," to discourage

EXAMPLE 3.1 INTERPRETATION 1—ONE FROM THE DANCE

We should always have in mind that our 95% CI is one from an infinite sequence of repeated experiments. Researchers rarely write about this basic way of interpreting CIs, but I can give one example that comes close. Scott, Lambie, Henwood, and Lamb (2006) reported an analysis of New Zealand crime statistics. Referring to a set of 96 convictions for rape, one question they examined was the relative chance that an intruder rapist (a person who intruded into a residence to commit the crime), compared with a nonintruder rapist, had a previous conviction for trespass. The ES they used was the odds ratio, which is one way of expressing relative chances, or relative risk. Their estimate was 5.91, 95% CI [1.72, 20.35], meaning that the odds for having such a previous conviction were about six times higher for intruder that nonintruder rapists. The authors wrote that the "odds of an intruder rapist [compared with a non-intruder] having a prior trespassing conviction lies 95% of the time, between 1.72 and 20.35" (p. 270).

When they referred to 95% of the time, they no doubt had in mind the definition of a CI as one from an infinite sequence of CIs from repeated experiments. Their wording, however, doesn't make that clear and may, I suspect, be confusing for many readers. If you wish to use Interpretation 1 in writing, it needs to be explained more fully. The key point is that the 95% refers to the whole process of taking a sample and calculating a CI, 95% of which will capture μ. Any particular CI, such as [1.72, 20.35], either does or does not capture μ, and so the 95% doesn't apply directly to that interval, but to the process that generated it. We should always keep in mind the first interpretation of a CI, but it may not provide the best approach for discussing data.

any misconception that μ is a variable. However, in my view it's acceptable to say, "We are 95% confident that our interval includes μ," provided that we keep in the back of our minds that we're referring to 95% of the intervals in the dance including μ, and 5% (the red ones) missing μ.

By saying I'm 95% confident that our CI contains μ, I'm saying that the values in the interval are *plausible* as true values for μ, and that values outside the interval are relatively implausible—although not impossible. We can thus consider substantive interpretation of the

Interpretation 2 of a CI. Our CI is a range of values that are *plausible* for the parameter we're estimating. The LL of our interval is a likely lower bound for the parameter, and the UL a likely upper bound.

various values in the interval. We'd probably first interpret the mean, at the center of the interval, which of course is just our point estimate. We could also consider the lower and upper limits of the interval. The LL is a likely lower bound for the true value of μ, although we know that just occasionally the LL won't be low enough (in 2.5% of cases, i.e., the red intervals that happen to land way to the right in the dance of the means). Similarly, the UL is a likely upper bound for μ (except in the 2.5% of cases in which it's not quite high enough—the red intervals that land way to the left). This interpretation of a CI as a range of values that are plausible for μ is probably the most widely used approach, and is often my favorite. For example, in discussing Tables 3.1 and 3.2 previously, I spoke of the 95% CI, which was [1.22, 6.95], as suggesting that the risk for the heaviest drinking group of mothers was raised by at least around 20% (referring to *LL* = 1.22) and perhaps by as much as a factor of about 7 (referring to *UL* = 6.95). Examples 3.2 use the second interpretation of CIs.

EXAMPLES 3.2 INTERPRETATION 2—THE INTERVAL AND ITS LOWER AND UPPER LIMITS

Example 3.1 came from Scott et al. (2006). Those researchers also used my second interpretation explicitly: "Confidence intervals are informative because they provide a range of plausible values" (p. 269).

Vaccination for Rubella

Sfikas, Greenhalgh, and Lewis (2007) reported a study of vaccination policies that could eliminate rubella from England and Wales. An important parameter is R_0, which is the average number of further infections produced by a single case of the disease. Sfikas et al. applied a somewhat complicated epidemiological model to a large database of blood samples to estimate R_0 = 3.66, [3.21, 4.36]. For any given value of R_0 they could apply their model and calculate the minimum proportion of children that must be vaccinated for the disease to be eliminated. The higher the value of R_0, the more infectious the disease, and so the nearer the vaccination rate must be to 100%. Assuming a single vaccination at birth, they calculated that the proportion of babies who must be vaccinated is .74, [.67, .76]. They commented that the point estimate is useful, but that the CI provides "a realistic idea of the limits within which the true proportion lies" (p. 6). Exactly. They go on to conclude that, in practice, "it may be more prudent ... to implement a campaign for which the

target vaccination proportion is closer to the upper 95 percentile limit rather than the point estimate in order to lower the risk of an epidemic" (p. 15). That's an example of interpretation of the CI as a range of plausible values of the parameter of interest, then a focus on one of the CI limits as a likely upper-bound estimate that should be adopted as the target for policy.

Does the Speed of Light Vary?

Abdo et al. (2009) is an article in the journal *Nature* with more than 200 authors from many countries. It reports astronomical data from ground stations and a telescope in orbit around the Earth that provide a test of an important challenge that has been made to Einstein's special theory of relativity. The theory postulates that the speed of light in a vacuum is always exactly the same, whereas the challenge suggests that quantum gravity effects might lead to variation in the speed of light over extremely small distances. Previous research had found evidence of invariance of the speed of light, as Einstein's theory predicts, down to 1.6×10^{-32} cm, which is an exceedingly short distance. Abdo et al. reported data that allowed them to push that boundary down even further. Their 99% CI was [1.6×10^{-35}, 1.3×10^{-33}]. They focused on the UL and claimed invariance of the speed of light has now been established down to 1.3×10^{-33} cm, a distance about one-twelfth the size of the previous boundary. They thus offered further support for Einstein's theory.

I don't claim to understand all aspects of Abdo et al. (2009), and my previous explanation is sketchy. You may be wondering why I chose this example. The researchers estimated the distance at which they had evidence of invariance of the speed of light and, naturally, they calculated a CI on their estimate. I'm interested because they chose to focus on the upper limit of their 99% CI. Their data gave reasonable evidence of invariance at even shorter distances, but they elected to choose the conservative end of their interval and claim they had shown invariance just down to 1.3×10^{-33} cm. Of course, the 99% CI they calculated for their data would be wider than the 95% CI, so the upper limit would be greater for the 99% than the 95% interval. By using the 99% CI, they were thus adding a further degree of conservatism.

Examples 3.2 prompt me to make two further comments. First, what confidence level should we choose? Some statisticians advise choosing a level to suit how concerned we are that our CI includes the parameter we're estimating. If it's a life and death matter, choose a 99% or even 99.9% CI to increase our confidence that our CI includes the true value. If we're not so concerned about the occasional miss, we might choose to report an 80% or 90% interval, these, of course, being considerably shorter than the 99%, let alone the 99.9% CI. That's reasonable advice, but, even so, my recommendation is to use 95% routinely, unless there are strong reasons for choosing some other value. I've seen so much evidence of misinterpretation of CIs—as some of the boxes throughout this book report—that I feel it's best to concentrate on understanding 95% CIs well, without the additional complexity of trying to interpret intervals with various different levels of confidence. However, in the next chapter we'll discuss how to translate easily between a 95% CI and an interval with some other level of confidence, so you should be able interpret a result, whatever level of confidence is reported. The Abdo et al. (2009) situation, in which the researchers chose one limit as the primary finding, is a case where it could be justifiable to use a 99% CI, or an interval with some other level of confidence you judge appropriate for the situation. The Sfikas et al. (2007) example of estimating the percentage of babies that need to be vaccinated is another case where the UL of a CI is used, and it may be prudent public policy to choose 99% or some other high level of confidence, rather than the 95% chosen by the authors. These are also cases in which we could consider a one-sided CI, rather than the usual two-sided CIs I've been discussing. One-sided CIs are not often used, but I discuss them in Chapter 4.

My second comment is that I hope you are getting the feeling that our approaches to the interpretation of a CI, two of which I've discussed so far, are very general. I'm deliberately choosing examples ranging over psychology, criminology, economics, ecology, medicine, astronomy, and other disciplines. Whatever the ES, whatever the situation, you can most likely use any of our six approaches to the interpretation of a CI.

CI Interpretation 3: The MOE Gives the Precision

Interpretation 3 of a CI. The MOE, which is the length of one arm of our CI, indicates *precision* and is the maximum likely error of estimation.

As we discussed, the MOE is the largest likely error of estimation, and so the MOE is a measure of the *precision* of our experiment. A third approach to CI interpretation is to use the MOE as indicating how close our point estimate is likely to be to μ, or the largest error we're likely to be making. It's easy to get tangled up in language about precision, because our measure of precision is MOE, but *increased* precision means

EXAMPLES 3.3 INTERPRETATION 3—THE MOE GIVES THE PRECISION

Scott et al. (2006) used this third interpretation of CIs by speaking of precision and width and the desirability of narrow intervals. However, they also commented that a wide CI can be "an indicator of uncertainty as to where the result falls" (p. 269), but that may be misleading. We know exactly where our result falls—for example, our *M*. The uncertainty is about the population parameter, and so it may be clearer to say something like, "uncertainty as to where the true value lies." Perhaps they meant "uncertainty as to where the result falls in relation to the unknown parameter." They also made a reasonable comment about the precision of one of their intervals by describing it as "quite broad because of the small sample size" (p. 270). They thus referred to precision and CI width, but not specifically to MOE. The next example focuses on MOE.

The Ages of Rocks

Broken Hill has long been an important mining city in outback Australia, and so there has been intensive study of the complex geology of surrounding areas. Rutherford, Hand, and Barovich (2007) reported estimates of the ages of a number of rock types, based on a large set of chemical analyses of rock samples. Their purpose was to use the age results to evaluate various models of how tectonic plates had moved and interacted in the area around 1.6 billion years ago. They reported their age estimates of the different rocks as, for example, 1594 ± 17 Ma (million years) and 1585 ± 31 Ma, where the ±17 and ±31 were stated to be 95% CIs, and so 17 and 31 were the MOEs. They reported and attended to MOEs throughout the article, for example, by commenting that "errors on mean ages range between 15 and 40 Ma" (p. 70). The precision of age estimates was important for their main conclusion, which was that an important tectonic event occurred between 1585 and 1610 million years ago, but that a previous suggestion of an earlier tectonic event occurring around 1690 million years ago was mistaken.

shorter MOE, and an *increase* in the MOE (taking a smaller sample, for xample) means *lower* precision.

The 1594 ± 17 format used by Rutherford et al. (2007) (see Examples 3) for reporting error margins is common in some disciplines, but it's ssential to be sure what quantity is being reported. Those researchers

stated explicitly that they were reporting 95% CIs, but the ± format is probably more often used to report the SE or SD. Further, in some disciplines including physics and chemistry it's common to use ± to report measurement error, but without any statistical definition. A report of, for example, 32.5 ± 0.1 mm suggests that the researcher simply judged that the length scale could be read to an accuracy of about 0.1 mm. No statistical definition was intended.

This is a good point to mention the article by Cumming and Finch (2005, tinyurl.com/inferencebyeye), which introduces and explains CIs, and describes a number of ways to interpret them. Many of the issues discussed in this chapter and in the following three chapters are also discussed in that article.

Exercises

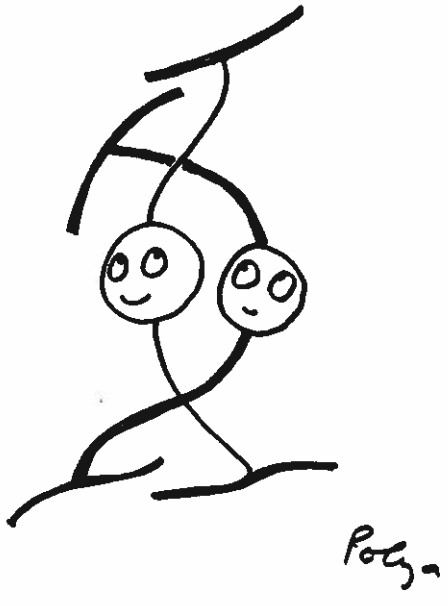
3.55 In your own discipline find, or compose for yourself, some examples of interpretation of CIs. Then for each make a second interpretation, based on some different approach to thinking about CIs.

3.56 Identify in each case which of my first three ways to interpret a CI is being used—or perhaps some other way is being used, or a mixture.

3.57 Think of the dance of the means, the mean heap, and the dance of the CIs. Are you dreaming about them yet? You are sufficiently familiar with them when they come up in your dreams.

3.58 Write down your own take-home messages from this chapter.

Reward yourself with chocolate or another play with ESCI if you actually write down your own before turning over to see mine.



Take-Home Messages

- Our statistical model assumes random sampling from a normally distributed population that has mean μ and standard deviation σ, which are fixed but unknown parameters.
- The simplest experiment is to take a single sample of size N from the population, and use sample mean M and sample standard deviation s as estimates of μ and σ.
- Sampling variability is the variability from sample to sample, and is illustrated by the dance of the means: Larger variability gives a wider, or more drunken, dance.
- *Take-home movie*: The dance of the means, as in Figure 3.2.
- The mean heap is the empirical sampling distribution of the sample means. After a notionally infinite number of samples it becomes the theoretical sampling distribution—illustrated in ESCI by the sampling distribution curve.
- *Take-home picture*: The mean heap, as in Figures 3.3, 3.4, and 3.5.
- The sampling distribution of the sample mean is normally distributed, with mean μ and SD of $\sigma/\sqrt{N}$.

- The SD of the sampling distribution is called the standard error, which gives the chant, "The standard error is the standard deviation of the sampling distribution of the sample mean." Just think of the mean heap: The SE measures its spread. It's worth remembering SE = $\sigma/\sqrt{N}$. That's Equation (3.1).
- The green stripe interval in ESCI extends MOE (the margin of error) either side of μ, and includes 95% of sample means. MOE = 2SE approximately, or more exactly MOE = 1.96 × SE = 1.96 × $\sigma/\sqrt{N}$.
- The MOE is the largest likely error of estimation. For 95% of samples, M lands within MOE of μ, or in other words, $|M - \mu| <$ MOE. For 5% of samples the mean falls outside the MOE lines, in a tail of the mean heap.
- The 95% CI extends MOE on either side of M, so the 95% CI is [M – MOE, M + MOE] or [$M - 1.96 \times \sigma/\sqrt{N}$, $M + 1.96 \times \sigma/\sqrt{N}$]. That's Equation (3.2). Those CIs are all the same length, based on known σ.
- We usually drop the unrealistic assumption of known σ and use s as an estimate of σ. Then the 95% CI is [$M - t_{95}(N-1) \times s/\sqrt{N}$, $M + t_{95}(N-1) \times s/\sqrt{N}$], where $t_{95}(N-1)$ is a critical value of t, with $(N-1)$ degrees of freedom. That's Equation (3.3).
- CIs based on s and t vary in length from sample to sample. Smaller N gives greater variation from sample to sample. For very small samples, CI length gives a poor indication of uncertainty, so we shouldn't trust CI length for such samples.
- *Take-home movie*: The dance of the CIs, as in Figure 3.8.
- The level of confidence, C, is usually set to 95, but can be given other values. Larger C gives wider CIs.
- Researchers should, wherever possible, report CIs for any ES estimates they report, then should interpret the ESs and CIs.
- See a CI reported, and automatically think, "It might be red!" Think of a CI as one from a potentially infinite dance of the CIs, C% of which capture the population parameter being estimated. That's the first approach to interpreting a CI. We can be C% confident our CI includes μ. But it just might be red.
- The second way to interpret a CI is as a range of values that are plausible for μ. The LL is a likely lower bound for μ, and the UL a likely upper bound.
- The third approach to interpreting CIs is to consider MOE as the precision of estimation. MOE is the largest likely error of estimation, meaning that the point estimate is likely to be within MOE of the parameter μ.