





## Estimation Methods for Single Samples

1. Introduction	2
2. The one-sample $t$ tool: confidence intervals and hypothesis testing	3
2.1. Confidence Intervals	3
2.2. Hypothesis Tests	8
2.3. Assumptions for use of the $t$ -distribution	9
3. Is the one-sample $t$ -tool legitimate for $n = 5$ ?	11
4. Bootstrapping with a single sample	13
5. Estimating required sample size	14
Appendix: The statistical oddities of working with $n = 2$ .	19

### Special sections:

	<b>Core Concepts:</b> This chapter is focused on methods for means of measured variables from a single sample. Sections that are conceptually at the core of that discussion will be highlighted
	Some passages herein are for the <b>statistical aficionado</b> , and can be skipped by others.
	<b>Nerd alert...</b> There are some passage herein that are particularly nerdy, and can usually be skipped. The photo to the left will be your warning.
	<b>NPS Fire ecologist</b> This text is written in a general fashion, suitable for audiences of all ages. Occasionally there are bits that are specifically aimed at NPS FEs. They are highlighted by the photo to the left.

## Section One: Introduction

Science research does not often do much with a single sample; sure, you might summarize the data from a single sample (i.e. present the mean and its SE, or if you are in a high-class mood, a confident interval). But not so often do you test whether or not the population mean meets a certain level, or is at least some level, or no more than some level. In science research, one inevitably is seeking to establish relationships or treatment effects of one sort or another.

However, should you find yourself working as an ecologist<sup>1</sup> (or other sort of scientist) in a management setting<sup>2</sup>, you might well have a mandate to study closely estimation from a single sample.

Our point? For scientific research purposes, this chapter might not have much of a draw, but for managers, the content here might be quite useful. And so we introduce a management-directed example, compliments of C. Farris, Esquire.

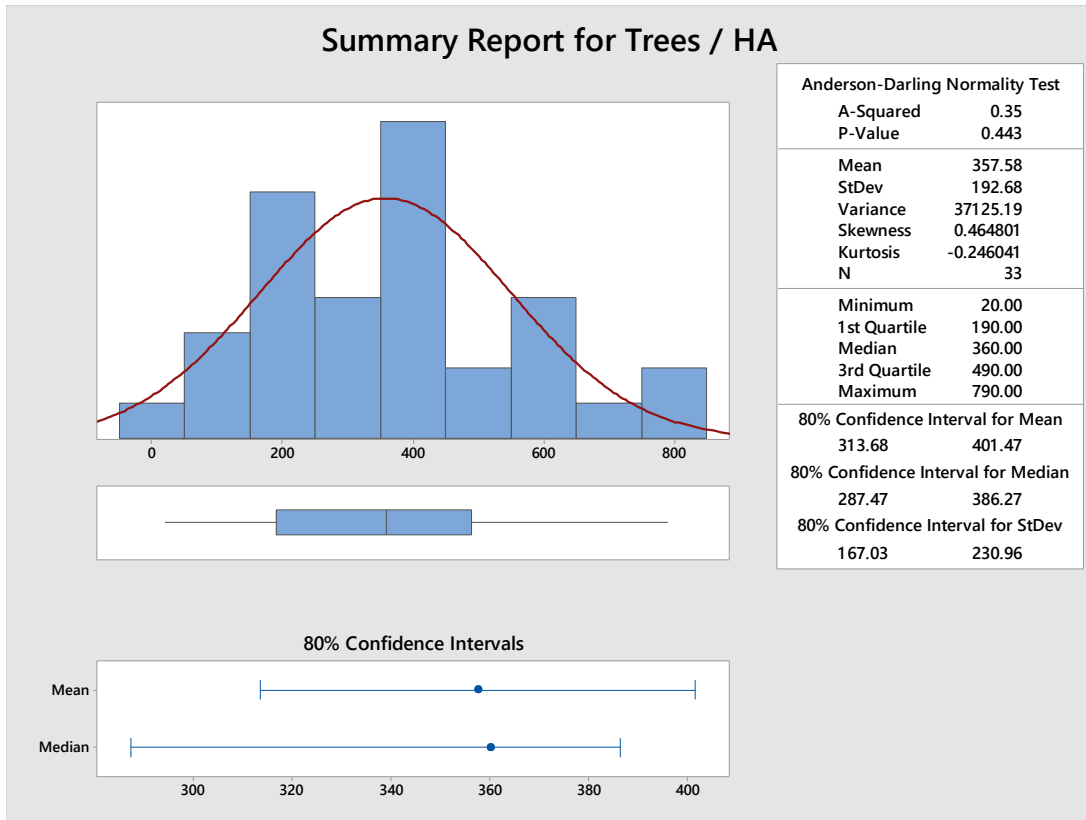
A controlled burn program in Lassen Volcanic National Park was instituted to reduce tree density (trees of all size classes) to within 50 to 300 trees per hectare (tph). Less than 300 would leave the landscape relatively safe from a catastrophic wildfire. At least 50 because they did not want to denude the landscape. Did they succeed? Figure 1 shows a graphical summary of the data (the managers chose a confidence level of 80%; we will come back to that later). It seems clear that they did not meet their objective, since they are pretty sure the average on the entire landscape is between 314 tph and 401 tph. We will address that formally later in this chapter, and also consider how to respond if, say, the objective had been to get between 100 tph and 400 tph. In the meantime, let's study tools to use on single samples when the goal is to estimate the mean (or test it against some chosen value).

---

<sup>1</sup> Or other sort of scientist

<sup>2</sup> For example, working as a government scientist, you might need to, in annual reports, show whether such and such has, on average, met a certain goal.

**Figure One.** Graphical summary of trees per hectare in burned plots at Lassen Volcanic National Park.



## Section Two. The one-sample $t$ tool: confidence intervals and hypothesis testing

### 2.1. Confidence Intervals



Some core ideas about confidence intervals and hypothesis tests are discussed here...

In this section, we will discuss making a confidence interval using the mean of tree density data; following that we will do a hypothesis test. For now, we will assume that the data are such that the one-sample  $t$  tool is valid<sup>3</sup>. Later we will introduce options should the  $t$  not be appropriate.

There are a number of things we need to unpack here. First is the meaning of a confidence interval. The choice of confidence level (e.g. 90% or 95%) is arbitrary<sup>4</sup>. The most commonly

<sup>3</sup> In fact, with  $n = 33$ , it's use is a safe bet, but we will come back to that.

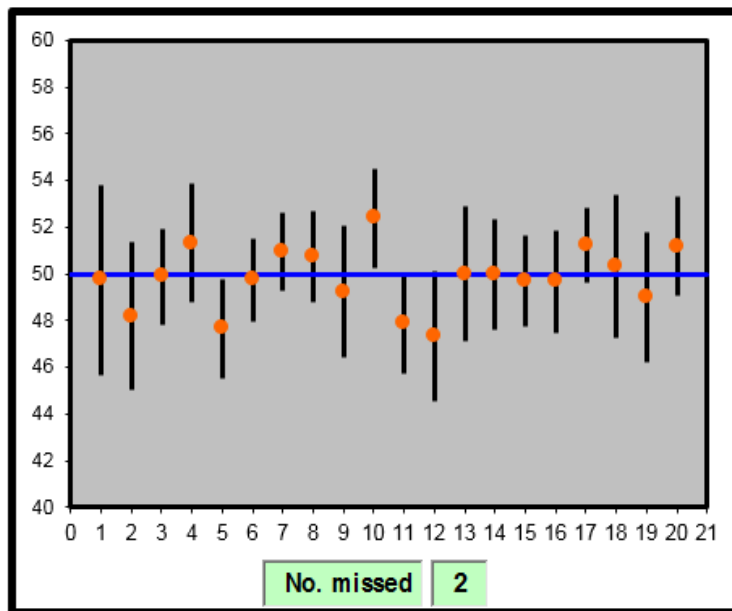
<sup>4</sup> Government agencies (e.g. state Game and Fish agencies, National Park Service) are often constrained to small sample sizes, and so sometimes choose to live with less confidence (e.g. 80%) in order to attain intervals that are narrow enough they can live with them.

used choice is 95%; it was first suggested by Ronald Fisher (he the inventor of ANOVA among other methods) almost 100 years ago, and has become a cultural convention ever since<sup>5</sup>. Here we will use 90% as our choice, both to be rebels, even if only in our own minds, and because it leads to a more visually compelling illustration.

Ken set up a computer simulation where data are generated from a Normal distribution with mean 50 (SD = 5). Sample size for this illustration was selected to be  $n = 10$ . Illustrated in Figure 2 are the confidence intervals from twenty such samples. Notice that 18 of the 20 actually include the true mean; 2 of them miss it<sup>6</sup>.

This illustrates the technical defining property of confidence intervals: if you repeatedly use the procedure, 90% (my choice here, for purpose of illustration) of the resulting intervals will in fact contain the parameter being estimate. This in turn is the foundation for the conventional statement: “I am 90% sure my interval contains the true parameter”. Maybe it does, maybe it doesn’t. All you have is your chosen level of confidence.

**Figure 2.** Depiction of twenty confidence intervals from a Normal distribution ( $\mu = 10; \sigma = 5$ ). Sample size for the simulations was  $n = 10$ ; chosen confidence level is 90%. The orange dots represent the sample means, while the vertical bars depict the intervals.



<sup>5</sup> The choice of confidence level is usually (and wisely) made to be complementary to the choice of alpha level (a.k.a. significance level) when doing a test. Ronald Fisher thought that a 5% chance of false significance when testing was a reasonable risk to take. It has since become almost dogma (which isn't good), but it is indeed often a reasonable choice. If alpha is 5%, then a confidence level of 95% is complementary.

<sup>6</sup> In this simulation, 18 (precisely 90%) of the intervals included the true mean; that is just luck. In repeats of this simulation, the number might be anywhere from 16 (rarely) up to 20. But *on average*, it will be 90%.

Now we unpack the construction construction of a 95%<sup>7</sup> confidence interval:

$$\bar{y} \pm t_{32,0.95} \times SE(\bar{y}) = 357.6 \pm 2.04 \times 33.5.$$

Here,

- (1)  $\bar{y}$  symbolizes the sample mean, used to estimate the (assumed fixed, but forever unknown) population mean, often symbolized by the Greek letter  $\mu$ .
- (2)  $t_{32,0.95}$  represents the value from a  $t$  distribution<sup>8</sup> with 32 degrees of freedom<sup>9</sup> such that  $\pm t$  captures the middle 95% of the distribution, and
- (3)  $SE(\bar{y}) = \frac{s}{\sqrt{n}}$ , is the estimate standard error<sup>10</sup> of the mean, where  $s$  is the sample standard deviation (SD; it estimates the population SD, often denoted by  $\sigma$ ). Here,  $s = 192.7$ , and  $n = 33$  is the sample size.

For these data, 95% CI would be (265.7, 449.5). As sample sizes get larger, the so-called  $t$ -multiplier  $t_{n-1,0.95}$  pretty quickly settles down to being quite close to “2”: For instance:

$$t_{9,0.95} = 2.26, t_{14,0.95} = 2.14, t_{19,0.95} = 2.09, \text{ and } t_{29,0.95} = 2.05.^{11}$$

With these same data, a 99% CI would be (289.3, 425.9), while an 80% interval is (313.7, 401.5). Notice that the 99% CI is wider than the 95% one, while the 80% one is narrower. The short version of this is that if you need to be surer that an interval captures the population mean, you must make it wider. If you are willing to be less sure, you get a narrower interval. Tradeoffs. Table 2 and Figure 2 illustrate the effect of sample size and choice of confidence level on the resulting  $t$  multipliers.

**Table 2.** The  $t$ -multipliers for 99%, 95%, and 80% confidence intervals for a variety of sample sizes (recall that degrees of freedom are sample size minus 1).

Confidence level	n = 5	n = 10	n = 15	n = 20	n = 30
99%	$t_{4,99} = 4.60$	$t_{9,99} = 3.25$	$t_{14,99} = 2.98$	$t_{19,99} = 2.86$	$t_{29,99} = 2.76$
95%	$t_{4,95} = 2.78$	$t_{9,95} = 2.26$	$t_{14,95} = 2.14$	$t_{19,95} = 2.09$	$t_{29,95} = 2.05$
80%	$t_{4,80} = 1.53$	$t_{9,80} = 1.38$	$t_{14,80} = 1.35$	$t_{19,80} = 1.33$	$t_{29,80} = 1.31$

<sup>7</sup> Most scientists routinely specify a 95% confidence level, so much so that stat packages use it as the default level. And so we will use it for the remainder of this explanation.

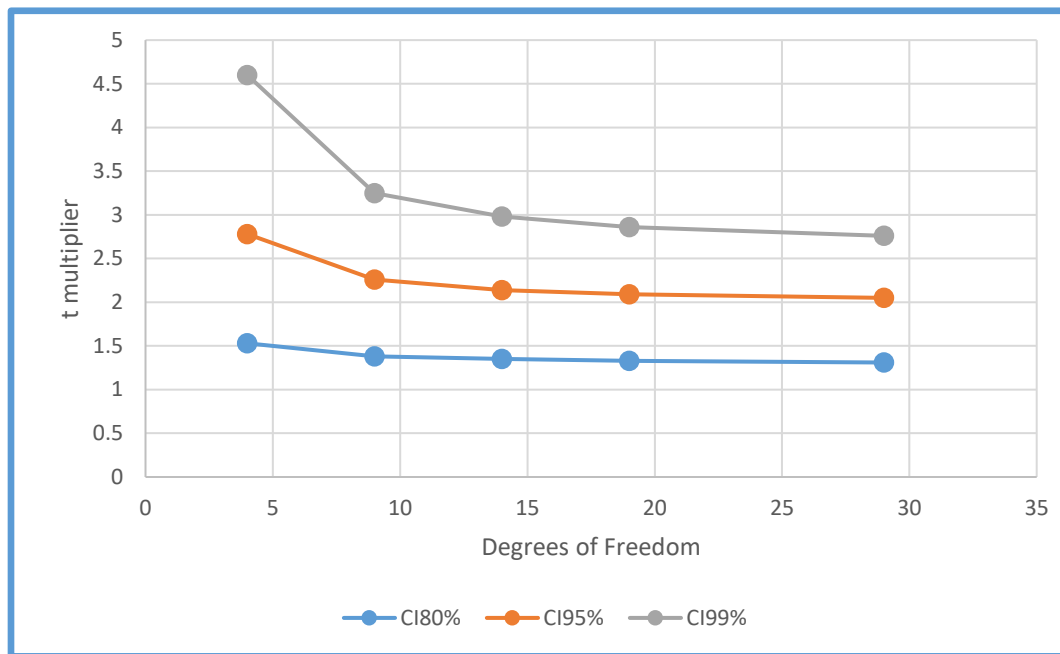
<sup>8</sup> We will explain below how and why a  $t$  distribution is used...

<sup>9</sup> The degrees of freedom formula for a single sample is just  $n - 1$ .

<sup>10</sup> We will go on record here to declare that “standard error” is an unfortunate piece of terminology. It is in fact an estimate of the SD of the distribution of the mean; we wish we could just call it that.

<sup>11</sup> This is the basis for a back of the envelope 95% CI that simply uses the number 2 for the multiplier. There is nothing wrong with creating a C.I. using “2” as a multiplier and calling it a PDS interval (PDS for “pretty darn sure”).

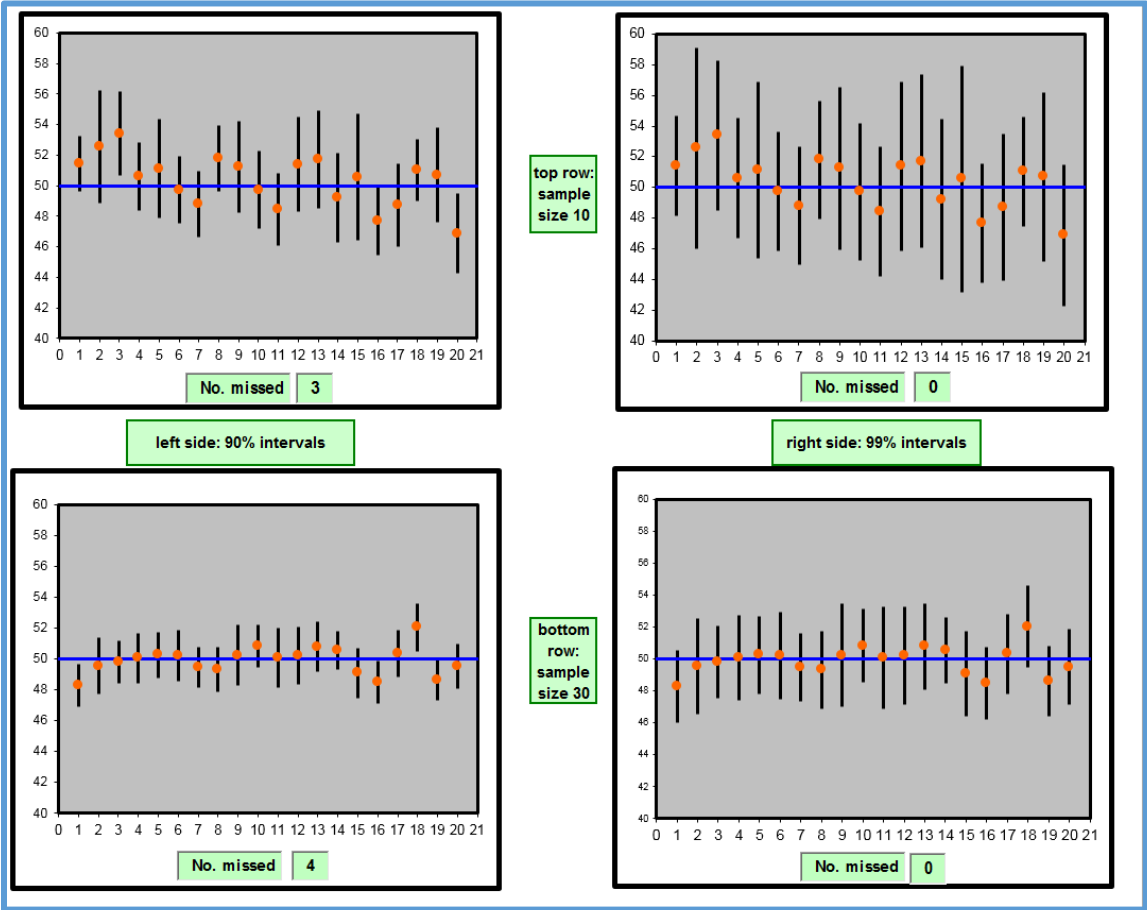
**Figure 2.** Graphical illustration (deploying data from Table 2) of how  $t$ -multipliers are affected by sample size and choice of confidence level.



Before we move on to more matters (specifically underlying assumptions and sample size considerations), let's take a moment to study CI behavior in a bit more detail. We summarize them in the following notes, as illustrated in Figure 3.

- (1) A 99% interval will be wider than a 90% interval. This might seem counterintuitive at first, since, after all, 99% implies “more sure”. But, with a given amount of information, if you want to be surer in making a prediction, you need to make it looser, wider. Imagine for instance predicting the number of points a sports team will score in a given game. If you want to be really, really sure your interval includes the actual result, a wider interval is called for. If you are willing to be less sure, you can use a narrower interval.
- (2) An increase in sample size will lead to narrower intervals. This does make sense intuitively: more data leads to more precise estimates. Pretty straightforward. But...
- (3) An increase in sample size does not lead to more “sureness”. That is purely and simply a reflection of choice of confidence level. A 90% interval will indeed be narrower if you have a sample of size  $n = 30$  instead of  $n = 10$ , but it will still lead you to being precisely 90% confident, no more, no less.

**Figure 3.** Simulations (each of size 20) of confidence intervals showing the effect of choice of confidence level (illustrated using 90% and 99%) and sample size (illustrated using  $n = 10$  and  $n = 30$ ).



## 2.2 Hypothesis tests

This section assumes passing familiarity with the principles and procedures for hypothesis testing. If you are rusty on them, please read the **Hypothesis Testing** Chapter in the **Big Ideas** section of the text.

In our tree density example, the ecologists in LVNP had set a goal of attaining a density of between 50 tph and 300 tph. The average from their sample was 357.6, so clearly, there is no evidence in favor of their attaining their goal. A 95% CI is (289.3, 425.9), so a value of 300 is not implausible. Yet...Let's see how the formalities play out.

Here, we need to (possibly) consider two tests, one taking aim at the upper targeted limit (300 tph), and the other at the lower (50 tph). The question of interest for the first test is whether the density could be less than 300 tph. This leads to a one-tailed alternate hypothesis:  $H_A: \mu < 300$ , which in turn determines the null:  $H_0: \mu \geq 300$ . Note that the null hypothesis<sup>12</sup> is indeed the nullity of the alternate hypothesis. There is no need to continue the test formally. Since the mean from the sample is *not* less than 300, the  $p$ -value will be greater than 0.5. There is no evidence in these data to support the contention. Given that, there is then no need to formally ask whether the mean is greater than 50. Clearly it is.

For sake of pedagogy, let's suppose the management criteria had been to get the density to between 200 tph and 450 tph. Now the situation is not so clear. The sample mean (357.6) clearly falls below the upper limit, but is it sufficiently far below that we can assert that it is unlikely to have come from a population whose mean is 450 tph or greater? So now we have :  $H_A: \mu < 450$ , leading to  $H_0: \mu \geq 450$ . The  $p$ -value for this one-tailed test is 0.005, clear evidence in favor of the alternate hypothesis<sup>13</sup>. For sake of continuing this conversation, let us accept that they are below the posited upper limit.

What about the lower limit? For this, we have  $H_A: \mu > 200$ , leading to  $H_0: \mu \leq 200$ . The  $p$ -value for this 0.000 (i.e. is less than 0.0005). Clearly they have met their management goal. Notice also that their 95% CI (289.3, 425.9) falls clearly within the prescribed range.

---

<sup>12</sup> The hypotheses are usually presented in the order of (null, alternate), but this makes little sense since the null is in fact derived from the alternate. You cannot know what will be the null until you specify the alternate.

<sup>13</sup> Tis so for most usual choices of alpha, from 0.01 and on up.

### 2.3. Assumptions for use of the $t$ -distribution

The foregoing use of the  $t$  for confidence intervals rests on the assumption that *the distribution of the mean* is at least approximately Normal, which does not require that the data themselves have a Normal distribution. This notion (attributing Normality to the distribution of the mean) is important enough and an explanation long enough that it is in its own chapter (Central Limit Theorem) in the **Big Ideas** section of this text. But before that (conceptually, at least), we need to think about the phrase, “distribution of a mean”.

What on earth can *that* mean? After all, we only have a single number for the mean. No distribution there! However, we do have a sample of data, and can look at its distribution via (say) a histogram. Further, we can imagine that the sample comes from some population that must also have a distribution of values.



ALL of statistical inference hinges on an understanding of the existence of a distribution of a statistic (the discussion here is made explicit by using a mean from a single sample as illustration).

If we were to flip a fair coin and ask you, “what is the chance it came up heads?”, most of you would answer, “50%” or “50-50” or the like. In so doing, you instinctively conjured up a very large number of such coin flips (of which approximately 50% would be heads) and applied your understanding of the behavior of those coin flips to answer to the current question.

This concept applies also to the sample mean<sup>14</sup>. Imagine repeating the experiment a very (very!) large number of times in the blink of an eye. Imagine that every time you do, all you do is write down the mean. When you are done, you will have a distribution of those means.

So... distribution. Then it must have some shape (i.e. one could imagine a histogram of it), it must have a mean, and, since it has variation, it also has a standard deviation (SD). The mean of that distribution, assuming the original sample to be a random sample from some population, is  $\mu$ , the population mean. The SD of the distribution of the mean is  $\sigma/\sqrt{n}$ , conveniently estimated by  $s/\sqrt{n}$ . (Recall that  $s$  is the symbol for the sample SD). The phrase “standard error of the mean” is in fact just a synonym for “SD of the distribution of the mean”. We truly wish that whoever first studied this business had instead simply called it the SD of the mean, and had never invented the term standard error.

By way of illustration, a distribution of data values and the subsequent distribution of the mean for that situation are illustrated in Figure 4. These data are distances (in km) between encounters by a tropical ecologist of army ant colonies in the tropical rain forest at La Selva, a prominent research station in Costa Rica. Sometimes the next encounter was swift (the smallest observed distances were on the order of 0.05 km (or 5 meters)). The largest distance was about 14 km. The

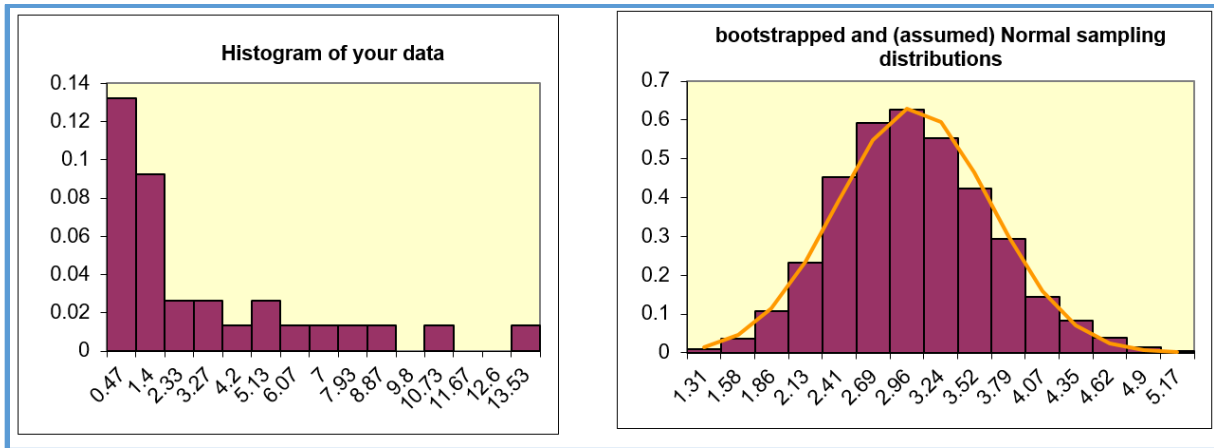
---

<sup>14</sup> Actually, the notion applies to *any* statistic (slope in a regression, difference in two means, a median, etcetera).

sample size was 25; the SD was 3.85. Clearly, the distribution of the data is *way* non-Normal. It is severely truncated at zero (as are many things we measure), and highly skewed to the right.

Yet bootstrapping<sup>15</sup> shows that the distribution of the mean is quite symmetric; one could declare it to be approximately Normal without offending anyone. The range of values in that distribution is quite smaller than in the raw data as is the SD (approximately 0.76) therein.

**Figure 4.** Illustration of the distribution of a mean (simulated by bootstrapping) from a sample of 25 distances between encounters of army ants on a research station in Costa Rica.



Ok. So, we can move ahead with the notion that a statistic (the sample mean here) has a distribution; that distribution has a mean and it has an SD. What about its shape? This is when things get really eerie: given a sufficiently large sample size, its shape will be well approximated by a Normal distribution, as illustrated in Figures 4.

This is a result of the statistical law of gravity called the Central Limit Theorem. How large a sample do you need for that result to obtain? Surprisingly, it is often not a very large number<sup>16</sup>, but learning to judge that is more involved than we want to tackle here. Older text books used to suggest that a sample size of 30 was required,<sup>17</sup> but the CLT often kicks in for quite smaller sample sizes. That advice from days of yore predated our ability to study this with computer simulations, and so the advice was conservative. Ken routinely refuses to say out loud the number “30” when teaching for fear that someone will write it down ☺.

At this point in your reading (our writing), we assume you have some idea of the notion of a distribution of a mean, and know that one can argue often that the distribution is approximately Normal. If this is so, why don’t we use the Normal distribution when making inferences on means? Where does this *t* distribution come into the picture? That is an interesting tale, but perhaps a side tale, and so we relegate it to a discussion in the Central Limit Theorem chapter in the **Big Ideas** section of this text.

<sup>15</sup> An introduction to bootstrapping is in this text, only pages away....

<sup>16</sup> But certainly not  $n = 5$ , like the example in these notes.

<sup>17</sup> With some variation; Calvin remember being told that 25 would be sufficient.

### Section Three: Is the one-sample $t$ -tool legitimate for $n = 5$ ?

The example we have used here has  $n = 33$ , and we are quite confident that use of the  $t$  is quite well-founded. But suppose you had a quite smaller sample size<sup>18</sup>? Suppose it was only 5?

The  $t$ -distribution is a technical variation on the Normal distribution: it is typically a bit wider (more so for smaller sample sizes), but it otherwise similar in shape: symmetric and bell-shaped. It is the correct distribution to use for a sample mean if you can argue that the distribution of the mean is approximately Normal. There are two bases that one could use for that argument.

**Basis 1.** If the random variable one is studying has a Normal distribution<sup>19, 20</sup>, then the distribution of the mean is Normal also, for any sample size.

The language we used above is similar to what is in most textbooks. Better, we think, would be to say, “If the population from which came our sample is Normal...” because that leads more naturally to using the histogram of the sample itself as a basis for estimating the distribution of that population.

**Basis 2.** If the sample size is large enough, the distribution of the mean will be approximately Normal. This is due to the Central Limit Theorem (CLT), which applies to many, but not all, commonly used statistics.

So we need to be able to examine our sample (the histogram of our sample, in particular) to see what we can learn from it about the distribution of the population. First point: with a small sample size, you cannot trust the histogram of the data to resemble the population whence it came. So: for small sample sizes, **testing for Normality is a waste of time.**



In fact, asking whether or not your data are Normal (and testing for it using your data) is such common practice that we feel obliged to show that it doesn't work for small sample sizes.

---

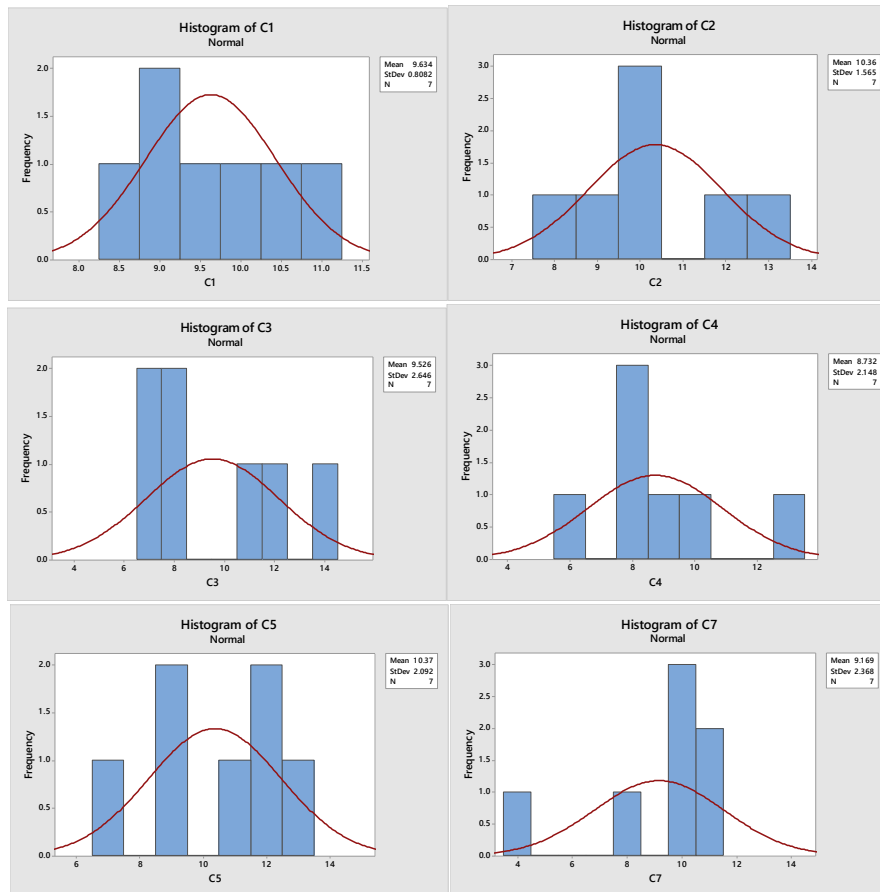
<sup>18</sup> This is unlikely to happen in a purely science research, but it not beyond imagining for (say) state or federal government folks, who often need to keep track of many variables, and may only be able afford a small effort on each.

<sup>19</sup> This is a pretty usual assumption given in text books. “If  $Y$  has a Normal distribution,...”. But it is pretty useless. First of all, most data we study *don't* have a Normal distribution. Second... see Basis 2 above.

<sup>20</sup> This argument can sometimes be legitimately made, but has to be based on information external to your data. For instance, if one has lots of familiarity with data like your current batch, **AND** one is confident from that experience that the distribution one is working with is approximately Normal (but see the preceding footnote), then that argument can render use of the  $t$ -distribution legitimate.

The histograms in Figure 5 are of randomly generated data ( $n = 7$ ; a Normal curve is drawn for comparison). What do you guess is the shape of the population whence they came<sup>21</sup>?

**Figure 5.** Examples of random samples of size  $n = 7$ , drawn from some mystery distribution. A Normal curve (using the mean and SD from the actual sample) are drawn on each histogram.



And what about large sample sizes? For a large enough sample size, we don't particularly care about the distribution whence came the data, because the CLT assures us that the distribution of the mean will be at least approximately Normal, which renders valid use of the  $t$  distribution. So: for large sample sizes, **testing for Normality is a waste of time.**

Ok, so we got a wee bit hyperbolic there, but our intention was/is to deflate the supposed importance of asking whether or not your data are Normal. Folks most often phrase the question as: "Are my data Normal?" That's a bit off-topic, because what you care about (if at all) is whether or not it appears that the population whence came your data is Normal; you are usually stuck with assessing that by looking at the histogram of your data.

<sup>21</sup> They came from a Normal distribution, with mean 10 and SD 2. Tough to tell, hunh?

Sometimes one has data sets that are sufficiently small ( $n = 10$  or less)<sup>22</sup> that they are too small for legitimate use of the  $t$ -distribution for hypothesis testing and/or confidence interval building (unless Basis 1 above can be deployed). What recourse, then? For making a confidence interval, bootstrapping is valid for *any* sample size bigger than  $n = \text{two}$ <sup>23</sup> (never mind that  $n = 2$  sucks). We turn now to an introductory treatment of bootstrapping.

#### Section Four: Bootstrapping with a single sample.



Bootstrapping is useful when (1) the statistic you are using doesn't fit into the classical canon of Normality-based methods<sup>24</sup>, or (2) your sample size is small, and so you cannot count on the Central Limit Theorem to work for you.

Any time you employ statistical inference, you are doing one of two things: generating a  $p$ -value from a hypothesis test of some parameter, or generating an estimate of some parameter. In this chapter, we are discussing the latter, with the estimate being accompanied by a confidence interval<sup>25</sup>. The  $t$ -tool approach depends on successfully arguing that said distribution is approximately Normal. For small samples, this argument is often weak or nonexistent. In that case bootstrapping can be used to construct a confidence interval. Bootstrapping for a mean from a single sample proceeds as follows. It is based on the idea that your data themselves are the best representation you have of the population whence they came.

**Assumption:** We assume the data are a suitable random sample from the population of interest.

#### Conceptual algorithm.

- (1) Write down each of your  $n$  sample values an infinite number of times<sup>26</sup>. This will generate a population (of infinite size) from which to do the simulation.
- (2) Draw a random sample of size  $n$  from that population. Write down the mean.

---

<sup>22</sup> Please don't take this to mean that 11 is good, and 10 is bad...The "10" here is a hand-wavy ballpark number. There *is* no sharp limit.

<sup>23</sup> Strictly speaking, one can do so for  $n = 2$  also, but the resulting intervals are pretty much useless, despite being validly constructed.

<sup>24</sup> An example might be the difference or ratio of two medians.

<sup>25</sup> It is entirely legal to report a mean and its SE (standard error) without making a confidence interval. Then the inference (estimate and a statement of precision) stand without requiring any assumptions about the shape of the distribution of the mean. **Cal editorializes:** This is interesting, even though the *real* reason we plot SE is because it is smaller than SD or CI (just being cynical).

<sup>26</sup> You might immediately think that this is impossible, and you would be correct. Hold your breath, and wait. The practical form of the algorithm provides a solution to this seemingly impossible step.

- (3) Repeat (2) a large number of times ( $B = 1000$  will often suffice,<sup>27</sup> although our Excel tools use  $B = 10,000$ , just because we can without inducing long waiting times<sup>28</sup>).
- (4) Sort the values from smallest to largest.

Then, for  $B = 1000$ , number 26 (isolating 2.5% that are smaller) and number 975 (isolating 2.5% that are larger) in that sorted list are the lower and upper bounds on a 95% confidence interval<sup>29</sup>. For an 80% interval, use numbers 101 and 900 (isolating the outer 20% and capturing the central 80%) from that list. Stunningly simple, except for step (1).

What step (1) implies, if you think about it a bit, is that any one datum in your original sample could appear one time, twice, thrice (less likely) or even zero times in any given bootstrap sample. How can we trick the computer into doing that without having to face infinity<sup>30</sup>? Read on.

### Practical Algorithm

Replace steps (1) and (2) above with: draw a random sample of size  $n$  from your original sample, *sampling with replacement*. The next two steps are the same.

What does “sampling with replacement” mean, and why does it work? First, the “what.” Column 4 of Table 1 has five values, namely the five differences (post-fire minus pre-fire) in vegetation cover values (the single sample being studied). The data are: 6, 12, 17, 26, and 18. Independently randomly select numbers from that list (independent selection allows a given number to be repeated in your list). For instance you could<sup>31</sup> get 12, 6, 18, 6, and 17, yielding an average of 11.8. The datum “6” appeared twice in the bootstrap sample.

Why does this work? By sampling with replacement, we perfectly mimic what step (2) of the conceptual algorithm does if you could actually write down all the values an infinity of times (step (1)). A simple computational trick replaces infinity! Gratitude to Brad Efron, the inventor of the bootstrap, for this ground-changing idea.

The bootstrapped and  $t$ -based 80%, 95%, and 99% confidence intervals for our leading example are in Table 3 below. For both methods, higher confidence levels require wider intervals, as previously discussed. The bootstrap intervals are distinctly different from the classical ones. This provides evidence that the distribution of the means is not particularly Normal, suggesting that

---

<sup>27</sup> The number of times you repeat this is the size  $B$  of the bootstrap simulation. It is important to know that should you choose  $B = 1000$  or  $B = 100,000$ , you are still studying inference for a sample of size  $n$ .

<sup>28</sup> There used to be arguments whether  $B = 100$  was enough or that one should use a larger number. Those arguments were germane when computing speeds were slow. When Ken first started creating bootstrap apps, he used 1,000, which cost a minute or so of waiting. Currently 10,000 (which is *way* plenty) only takes a blink of time.

<sup>29</sup> This is the so-called empirical bootstrap. There are more sophisticated approaches (a popular one being the so-called bias-corrected, accelerated (BCA) bootstrap method that employ some clever machinations on that simulated distribution. We won't be discussing those here; they are not necessary for the core ideas.

<sup>30</sup> We will have to face it eventually, but hopefully only after a reasonably long and good life.

<sup>31</sup> Ken *did* get precisely this using a random number generator.

the bootstrap intervals should be used. Luckily for us, they happen to be narrower as well, which is pleasing.

**Table 3.** Bootstrapped and *t*-based 80%, 95%, and 99% confidence intervals

Confidence level	Bootstrap	Classical ( <i>t</i> -based)
80%	(316, 402)	(314, 401)
95%	(294, 426)	(289, 426)
99%	(276, 448)	(266, 449)

**Practical tool:** The interactive Excel tool **one-sample boot** that accompanies this book can be used to do the calculations. **KG note to self: need to program the tool to allow confidence levels other than 95%..**

### Section Five: Sample Size Calculations for Confidence Intervals: what *n* do we need, anyway?



Sample size calculations are not employed by all, even though they are sometimes useful, especially if resources are limited. If you are just trying to learn the basics, you can skip this section for now.



The sample size formula below comes from page 216 of the Fire Monitoring Handbook (2003). That said, it is a pretty typical presentation of a sample size formula.

A commonly presented sample size formula is given as

$$n = \frac{t^2 s^2}{d^2} = \left( \frac{ts}{d} \right)^2,$$

where:

*n* is the sample size,

*t* is the required multiplier from the appropriate *t* distribution,

*s* is the sample standard deviation, and

*d* is the desired precision, given as a chosen percentage of the mean<sup>32</sup>.

<sup>32</sup> This last piece can be devilishly difficult to come up with. In our experience, if we ask a researcher how wide they want a CI to be, they are inclined to leave the consulting session to go get a drink. It helps somewhat to shift the question from absolute terms (i.e. in whatever are their measurement units) to *relative* terms. It is somewhat easier to land on, for instance, “OK. I want to get within 10% of the true value.”

If you have a pilot study, then the mean and SD can be estimated from it. Ideally, you would do a sample size calculation before launching on your study, in which case you do not yet have your data, in which case neither the SD nor the mean are known<sup>33</sup>. Further, as made clear in this chapter, the numerical value of  $t$  depends on both choice of confidence level *and* on sample size, since degrees of freedom =  $n - 1$ , and the  $t$  distribution is indexed by that number. So you cannot use the formula unless you know the sample size<sup>34</sup>...

All is not lost; we can unpack this and come up with an answer<sup>35</sup>. We are going to use the data in our lead example as a rough guide, but only that: a rough guide... Let's suppose you want to make an 80% confidence interval for the mean (common enough among NPS ecologists).

- (1) Precision: it is pretty natural to state precision in percentage terms, something like, "I want to get within 10% of the truth." (Truth here being the parameter being estimated). You need to, then, ahead of time, come up with a guesstimate for the mean.

**Example:** You guess that the mean tree density will be about 400 tph, and you declare that you want to get within 10% of the truth. That implies a confidence interval half-width<sup>36</sup> of  $.10 \times 400 = 40$ . This will be  $d$  in the formula.

- (2) Standard deviation  $s$ : this is incredibly difficult to guess at, but another measure of variation, the range of data values (largest minus smallest) is usually quite easier, and is related to  $s$ . It turns out that the range of values in most data sets is somewhere near 3 or 4 times the standard deviation. For smaller sample sizes, that number is closer to 3; for larger, 4.

**Example:** Suppose you guess that the smallest observed density will be nearly zero, and the largest almost 1000. A range of 1000 divided by four suggests a guesstimated<sup>37</sup> SD of 250.

- (3) Now we could go apply the formula, except for that darn  $t$ . Here's an algorithmic workaround (meaning there are a couple of steps to it).

First, **replace  $t$  with the analogous number  $z$** <sup>38</sup> from the standard Normal distribution:

---

<sup>33</sup> That alone could leave you inclined to give up and go get a pedicure instead.

<sup>34</sup> Commence eyeball spinning...

<sup>35</sup> What we are about to show you has several steps and more than a little tedious arithmetic. Luckily for you, Ken has created an interactive Excel sheet (called **SSCI onemean**) to do that work for you. So read through this, but then go use the tool.

<sup>36</sup> This is the +/- part of the CI formula:  $t \times SE$ .

<sup>37</sup> When you do this for real, we suggest running the numbers with these initial guesses, and then again with more conservative guesses. Here, for instance, having come up with a guess of 7 for the SD, we might run the numbers again using, say 9, which is on the order of about 30% larger...

<sup>38</sup> Values from a Normal distribution do not depend on sample size. For examples (compare to Table 2 or Figure 1, where we showed  $t$  multipliers for confidence intervals),  $z = 2.58$  for a 99% CI, 1.96 for a 95% CI, and 1.28 for an 80% CI.

$$n = \left( \frac{zS}{d} \right)^2.$$

The  $z$  value for an 80% CI<sup>39</sup> is 1.28. Let's see how that works out. We would get

(1)  $n = \left( \frac{1.28 \times 250}{40} \right)^2 = (8)^2 = 64.$

(2) For  $n = 64$ , degrees of freedom are  $n - 1 = 63$ . The correct  $t$  from that distribution for an 80% confidence interval is 1.30.

(3) Also, with  $n = 64$ , the estimated SE of the mean will be

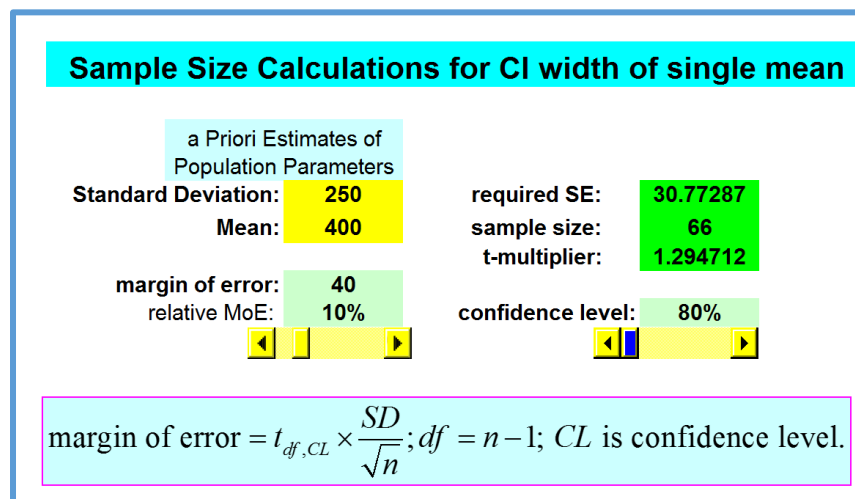
$$SE = \frac{250}{\sqrt{64}} = 31.25. \text{ Let's give it a go.}$$

(4) We get for the half-width  $t \times SE = 1.3 \times 31.25 = 40.6$ .

(5) This is *slightly* wider than our stated goal; bumping  $n$  up to 66 gets us there<sup>40</sup>.

Somewhat tedious, as we said, but the Excel calculator takes the pain out of that (screenshot in Figure 6). But this (realistic) example raises another, even more problematic point: you will almost surely not be able to afford 66 plots! Crap! What to do<sup>41</sup>? Lowering the confidence level will enable us to achieve our precision goal with a smaller sample size, but (darn it!) we are already committed to a relatively low level of confidence.

**Figure 6.** Screenshot of Excel calculator with data values as per our sample size example.



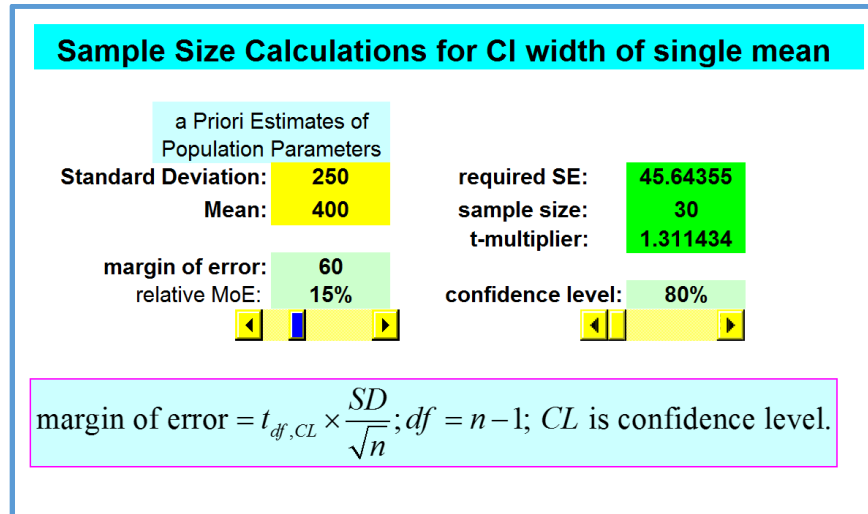
<sup>39</sup> For quick reference, the relevant value for a 95% interval is 1.96, while for a 99% interval, it is 2.58.

<sup>40</sup> **It will usually be the case that the initial value (determined from the  $z$  distribution) will be too low by one or two.**

<sup>41</sup> We warned you there would be very many footnotes; this one, in particular, is quite useless.

If indeed, in your situation, a sample of size 66 is not feasible, you could choose to soften your precision requirements<sup>42</sup>. For instance, the next screenshot (Figure 7) shows what the calculations would look like if you declared a willingness to get within 15% of the true value, instead of 10%.

**Figure 7.** Screenshot of Excel calculator with data values as per our sample size example, but with the precision requirements relaxed.



It appears that we can get to within 15% of “truth” with a sample size of about 30.

### Closing Thoughts.

So, for a single sample, with intent to estimate the mean, we have covered the role of the Central Limit Theorem, which makes legitimate the use of a  $t$  distribution, if one can argue for the distribution of the mean of differences being approximately Normal. This argument can be made by appealing to large sample sizes; failing that, one must argue from bases external to your data. If the  $t$  distribution is not applicable<sup>43</sup>, bootstrapping is sure to be legitimate for producing confidence intervals.

<sup>42</sup> Reducing the confidence level will also reduce the sample size requirements, but here we are sitting at a choice of 80% which is on the low side already.

<sup>43</sup> Which will be usual for NPS fire ecology data sets, they tending to be on the small side...

## Appendix: The statistical oddities of working with $n = 2$ .



**Nerd alert...** (this entire appendix)

Technically speaking, a  $t$ -based or a bootstrap confidence interval can work with  $n = 2$ . That said, let's investigate how (badly) that works.

### $t$ -based interval.

First, with  $n = 2$ , there is zero hope of the central limit theorem creating for you a Normal distribution for the mean; further, there is zero hope of ascertaining from your data whether or not they came from a population that has a Normal distribution. If you can argue for the latter for reasons external to your own data<sup>44</sup>, then the  $t$  is justified. **But.**

In **The Paired  $t$  tool** (Section 2 of this chapter), We showed you that the classical construction of a confidence interval has the form  $\bar{y} \pm t_{df, CL} \times SE(\bar{y})$ . Here,  $df$  is  $n - 1$ , and  $CL$  references the confidence level. There are different ways to denote that formally; no need to go into them here. The point is that the  $t$  multiplier changes depending on sample size and choice of confidence level in predictable ways. Further, the SE of the mean can get quite large for small sample sizes; both of these behaviors are illustrated in Table C.1, where we used  $s = 8$  (close to the actual standard deviation (7.43) from our veg cover data) for illustration.

**Table A.1.** Illustrative  $t$ -multipliers for a selection of sample sizes and confidence levels (this is a reprise of a table in the body of the chapter, but more detailed for purposes herein).

Sample Size	Confidence Level			SE ( $s = 8$ )
	80%	95%	99%	
2	3.08	12.71	63.66	5.66
3	1.89	4.30	9.92	4.62
4	1.64	3.18	5.84	4.0
5	1.53	2.78	4.60	3.58
10	1.38	2.26	3.25	2.53
20	1.33	2.09	2.86	1.79
30	1.31	2.05	2.76	1.46
50	1.30	2.01	2.68	1.13
$\infty$	1.28	1.96	2.58	0

- (1) For a chosen confidence level, as sample size goes up, it shrinks, quickly at first, and then ever-more slowly, eventually converging to the analogous value from a standard Normal (so-called  $z$ ) distribution.

<sup>44</sup> For example, you might have had scads of experience with data like yours, and can point to other, larger studies that could support the assertion, "these data come from a Normal distribution."

- (2) For a given sample size, as choice of confidence level increases, the multiplier becomes larger.

Even if one could justify it, confidence intervals for samples of size two are ridiculously wide. To illustrate, suppose  $\bar{y} = 16$  (close to 15.8, the actual estimated average change in cover from our working example).

An 80% CI for the change in cover would be (-19.7, 51.7), while a 95% CI would end up as (-55.9, 87.9). We won't even show you the 99% interval. In other words, the 80% interval is already so wide as to be useless.

### **Bootstrapping for $n = 2$ .**

Bootstrapping, while valid because it doesn't depend on an assumption<sup>45</sup> that can't be defended, is still untenable. It is easier to illustrate that with cartoon data, which we now do. Suppose you have a sample of size two, with observed values 3 and 5. The bootstrapping algorithm (see Section 4 if you need reminding) selects a sample of size 2, with replacement, from these two numbers. It is easy to determine that

- (1) 25% of the time, the resulting bootstrap sample will be (3, 3) with mean (of course) equal to 3;
- (2) 50% of the time one would get (3, 5), with mean 4, and
- (3) the final 25% would be composed of pairs of only the last data value: (5, 5), with mean 5.

So, if you did 1000 bootstraps, you would get for the distribution of the means approximately 250 3s, 500 4s, and 250 5s. So a confidence interval (of pretty much any chosen confidence level) would have as its lower limit "3", and "5" as the upper. True for a 99% interval, and for an 80% interval. Not very appealing behavior...

---

<sup>45</sup> Of Normality for the distribution of the mean