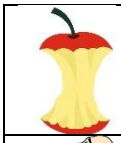
Estimation Methods for Two Independent Samples

Special sections:

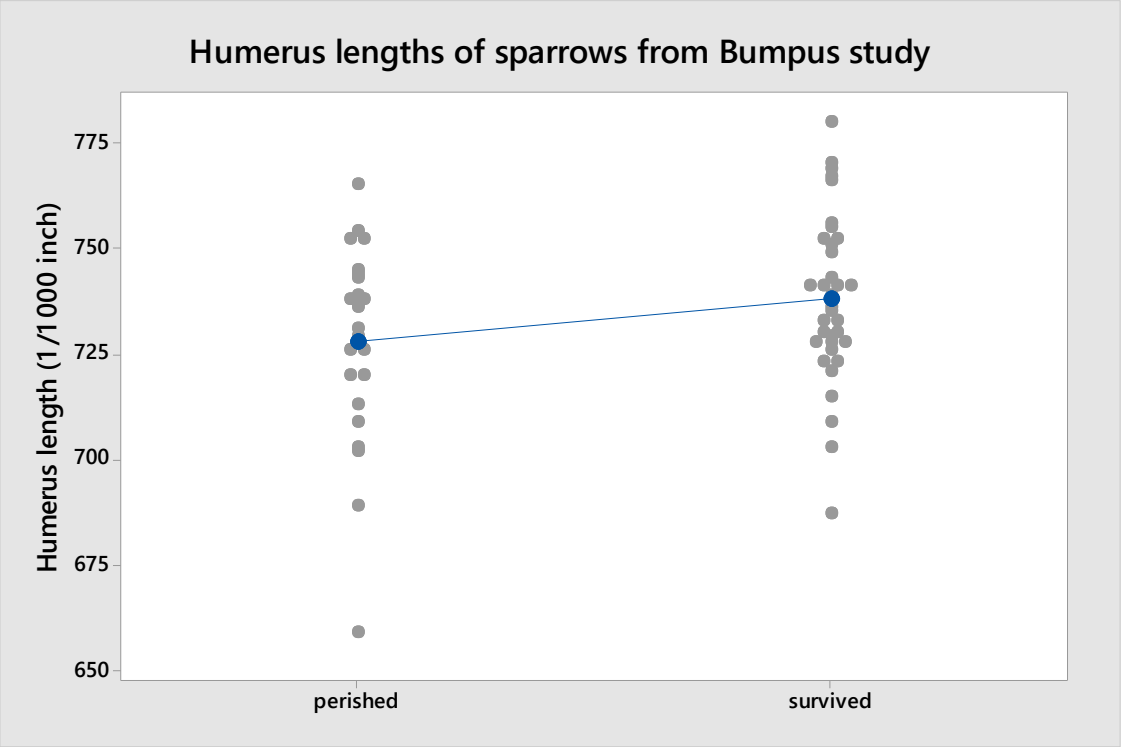
	**Core Concepts:** This chapter is focused on methods for inference on means from two independent samples. Sections that are conceptually at the core of that discussion will be highlighted.
	Some passages herein are for the **statistical aficionado**, and can be skipped by others.
	**Nerd alert…** There are some passage herein that are particularly nerdy, and can usually be skipped. The photo to the left will be your warning.
	**NPS Fire ecologist** 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

A common study design to examine the effect of a particular phenomenon (a certain treatment, or perhaps a given circumstance) is to have two independent[1] samples, one associated with the "treatment"; the other group is often called the control[2] group.

After a particularly severe winter storm in 1898, Hermon Bumpus collected (among other things[3]) measurements of house sparrow size, as measured by humerus length) for males[4] that perished in the storm (n = 24) and some that did not (n = 35); data displayed in Figure 1. Here we are interested in the question, "Did the survivors differ in size from those that perished?"

Figure One. Individual value plot[5] of sparrow lengths. Means are blue dots.



[1] A stronger design, when you can make it happen, is to have paired data. The fact that, within each pair, the two subjects are quite similar to start with, makes for easier detection of any effects. Technically, this is due to a smaller SE for the mean of the differences, caused by within-pair correlation of the data values.
[2] These designations are not mandated. For instance, if you wanted to compare males to females for some characteristic, the terms "treatment" and "control" would not apply.
[3] He actually measured nine characteristics on a total of 163 birds. For an interesting review of his study and its place in evolutionary science, see Johnston *et al*. 1971. Hermon Bumpus and Natural Selection in the House Sparrow *Passer domesticus*. Evolution 26:1, pp. 20 – 31.
[4] In house sparrows, males are slightly larger, on average, than females.
[5] Values that are repeats are spread out horizontally.

We will used these data (briefly summarized in Table 1) to illustrate the use of a two-sample *t* for estimation (via confidence intervals) and testing. Here interest is in the parameter[6] $D = \mu_S - \mu_P$, as estimated by $\hat{D} = \bar{y}_S - \bar{y}_P$.

Table 1. A brief numerical summary of the sparrow data

Variable	*n*	Mean	SE(M)	SD	Min.	Q1	Median	Q3	Max.
Perished	24	727.92	4.81	23.54	659	714.75	733.5	743.75	765
Survived	35	738.00	3.35	19.84	687	728.00	736.0	752.00	780

Section Two. The two-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 difference in mean lengths for the sparrow data; following that we will do a hypothesis test. For now, we will assume that the data are such that the two-sample *t* tool is valid[7]. Later we will introduce options should the *t* be not 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[8]. The most commonly 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[9]. 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

[6] In establishing the parameter of interest, choice of order of operation (perished minus survived *versus* survived minus perished) is arbitrary. We recommend ordering them so as to yield a positive number. It makes no matta to the analysis, but will make the story telling more convenient. Minus signs are confusing (to Ken, at least).

[7] In fact, with *n* = 24+35 = 59, it's use is a safe bet, but we will come back to that.

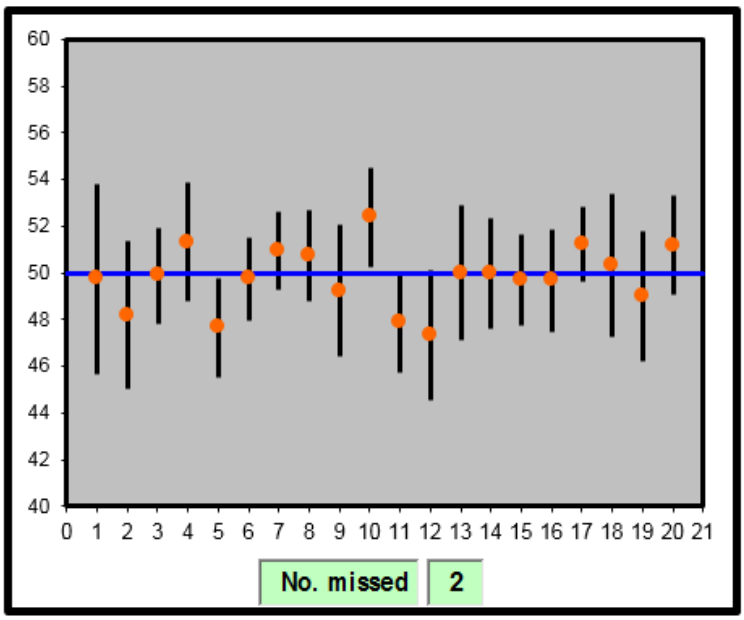
[8] Government agencies (e.g. state Game and Fish agencies, National Park Service) are often constrained to small sample sizes, and so sometimes choose to live with less confidence (e.g. 80%) in order to attain intervals that are narrow enough they can live with them.

[9] 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.

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[10].

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.



Now we unpack the construction construction of a 95%[11] confidence interval:

$$\hat{D} \pm t_{43,0.95} \times SE\left(\hat{D}\right) = 10.08 \pm 2.02 \times 5.85.$$

Here,

(1) $\hat{D} = \bar{y}_S - \bar{y}_P$ symbolizes the estimated difference in means.

[10] 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%.

[11] 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. We are over our rebel phase.

(2) $t_{43,0.95}$ represents the value from a *t* distribution[12] with 43 degrees of freedom[13] such that $\pm\, t$ captures the middle 95% of the distribution, and

(3) $SE\left(\hat{D}\right) = \sqrt{\left(SE\left(\bar{y}_S\right)\right)^2 + \left(SE\left(\bar{y}_P\right)\right)^2}$ is the estimated standard error[14] of the difference, the formula for each element is the formula is $SE\left(\bar{y}\right) = s/\sqrt{n}$, where *s* is the sample standard deviation (SD; it estimates the population SD, often denoted by σ). Here, $s_P = 23.5, n_P = 24, s_S = 19.8,$ and $n_S = 35.$

For these data, 95% CI for the difference is (-1.73, 21.9).

With these same data, a 99% CI would be (-5.71, 25.88), while an 80% interval is (2.46, 17.71). 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 3 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[15] (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%	$t_{4,80} = 1.53$	$t_{9,80} = 1.38$	$t_{14,80} = 1.35$	$t_{19,80} = 1.33$	$t_{29,80} = 1.31$

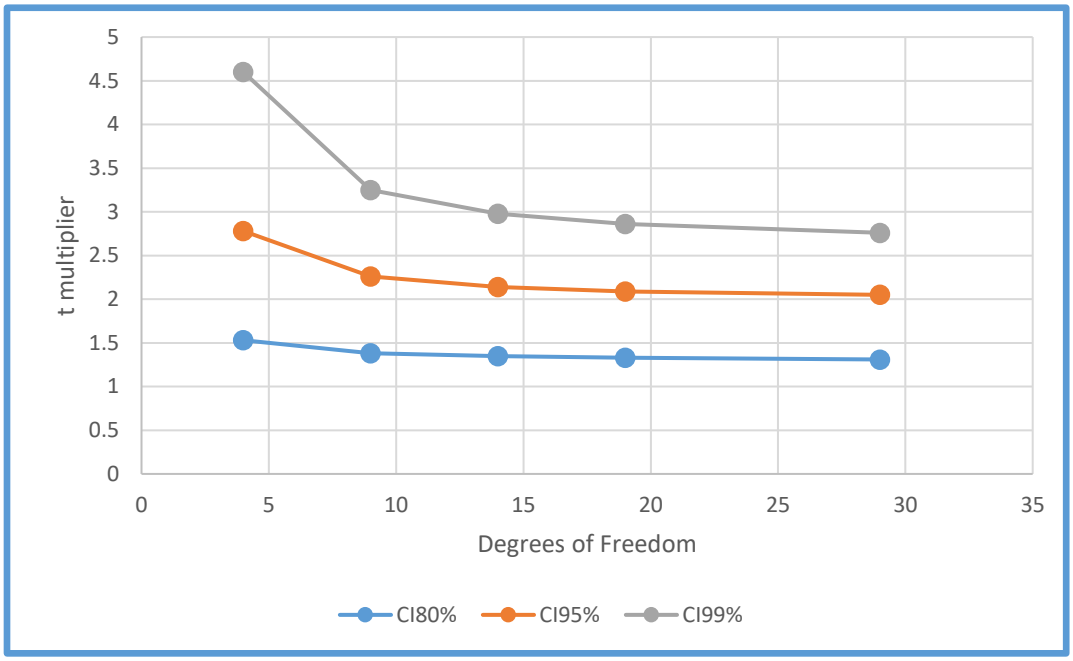
[12] We will explain below how and why a *t* distribution is used…

[13] The degrees of freedom formula for this case is a complicated formula, which we will explain shortly.

[14] 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.

[15] For simplicity here, we illustrate for the mean from a single sample.

Figure 3. Graphical illustration (deploying data from Table 2) of how *t*-multipliers are affected by sample size and choice of confidence level.



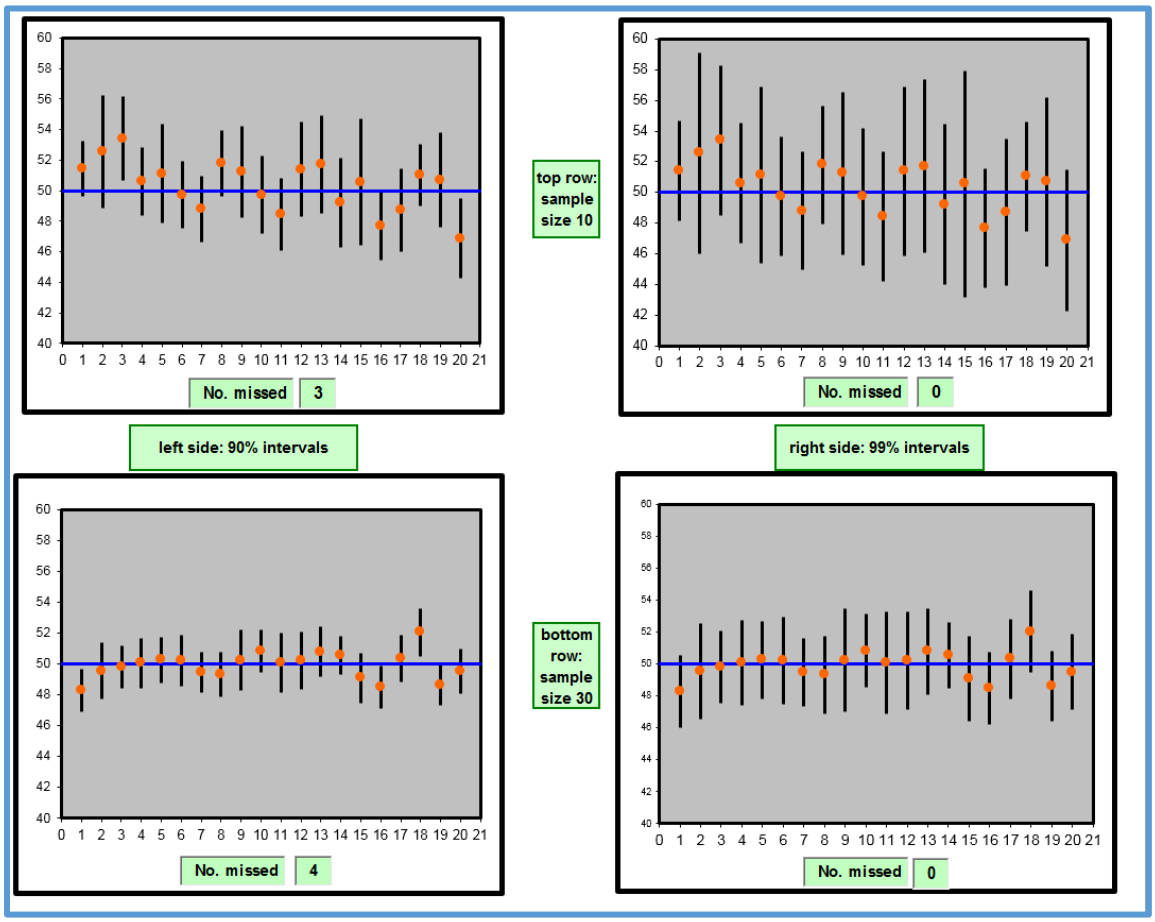
Before we move on to more matters (specifically underlying assumptions and sample size considerations), let's take a moment to study CI behavior in a bit more detail. We summarize them in the following notes, as illustrated in Figure 4.

(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 4. 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).



Important technical issue: equal variances or unequal?

When you use a two-sample t distribution to make a confidence interval (or do a test, more on which later), you need to choose whether or not to assume that the two samples come from populations with equal variances. For purposes of estimating the difference between the two means, we prefer to *not* assume equal variances, by the following argument. The phrase, "estimating the difference" asserts that there *is* a difference in the population means. If so, then the two samples come from different populations, each with their own variances. True, the two variances might be similar, but we see no need to insist that they must be equal. We see it as an un-necessary assumption. Our view on the matter for testing, however, is different.[16]

That said, it is certainly the case that there is no consensus on the matter, which we document in Appendix B.

[16] Patience. The hypothesis testing section is coming up.

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 test.

In our sparrow example, the question was posed as: do the two types of sparrows (survivors and casualties) differ in size, as measured by humerus length? Here, then, we have **H_A:** $\mu_S - \mu_P \neq 0$, leading to **H_0:** $\mu_S - \mu_P = 0$.

```
T-Test of difference = 0 (vs ≠):
T-Value = 1.78  P-Value = 0.081  DF = 57
Test used Pooled StDev = 21.4107
```

The *p*-value is larger than the conventional $\alpha = 0.05$, and so we would fail to reject the null hypothesis. Notice that we did the test using the "equal variances" condition. We do so because a hypothesis test begins with our standing firmly on the null, and saying, "show us" why we should reject it. In this case, we are formally testing for a difference in means, but the deeper implication is that the two samples do not differ with regard to size; that is, that they actually came from a single population. In that case, the two sample variances differ only by random chance. In short, asserting equal variances is consistent with the null hypothesis. See Appendix B for a deeper discussion of this.

In this case, we might be able to argue, as a research hypothesis, that larger birds might have a better chance of withstanding the hardship of a severe winter storm. In that case, we have **H_A:** $\mu_S - \mu_P > 0$, leading to **H_0:** $\mu_S - \mu_P \leq 0$. Since the actual difference (10.08) is consistent with the alternate hypothesis[17], we can simply take the *p*-value from the two-sided test and halve it: *p* = 0.04. Modestly significant against $\alpha = 0.05$.

[17] We note that had the alternate hypothesis been such that survivors were expected to be *smaller* than those that perished, the observed (sample mean of survivors is larger) would have stopped test, dead in it's tracks: there is clearly no evidence in the data in favor of the research hypothesis.

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 differences in means* 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 each 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[18]. 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 μ, 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.

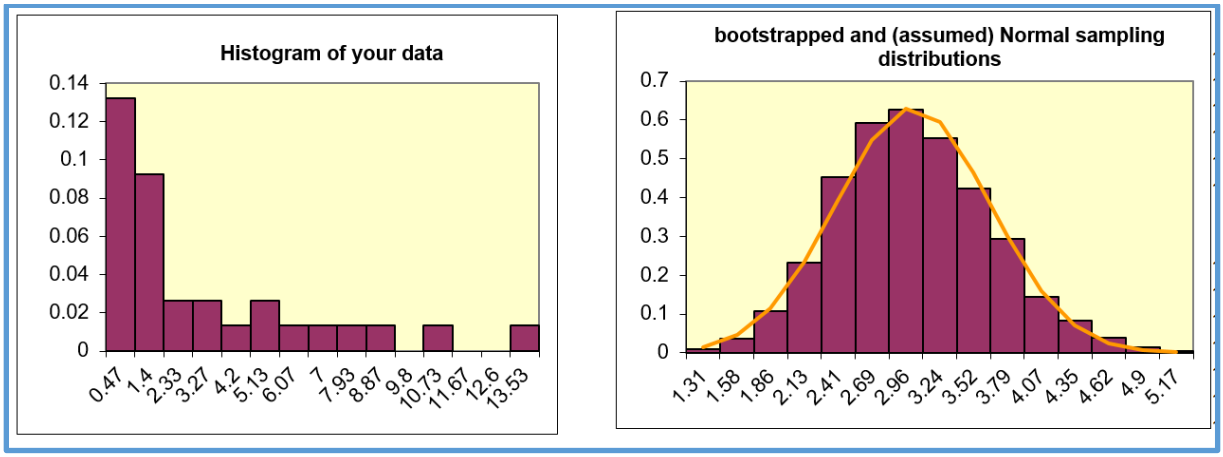
By way of illustration, a distribution of data values and the subsequent distribution of the mean for that situation are illustrated in Figure 5. These data are distances (in km) between encounters by a tropical ecologist of army ant colonies in the tropical rain forest at La Selva, a prominent research 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

[18] 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[19] 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 5. 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[20], 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,[21] 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.

[19] An introduction to bootstrapping is in this text, only pages away….
[20] But certainly not $n = 5$, like the example in these notes.
[21] 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 a total sample size of $n = 59$, and we are quite confident that use of the *t* is quite well-founded. But suppose you had a quite smaller sample size[22]? Suppose it was only 11, with, say, 5 in one group and 6 in the other?

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[23, 24], 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.

---

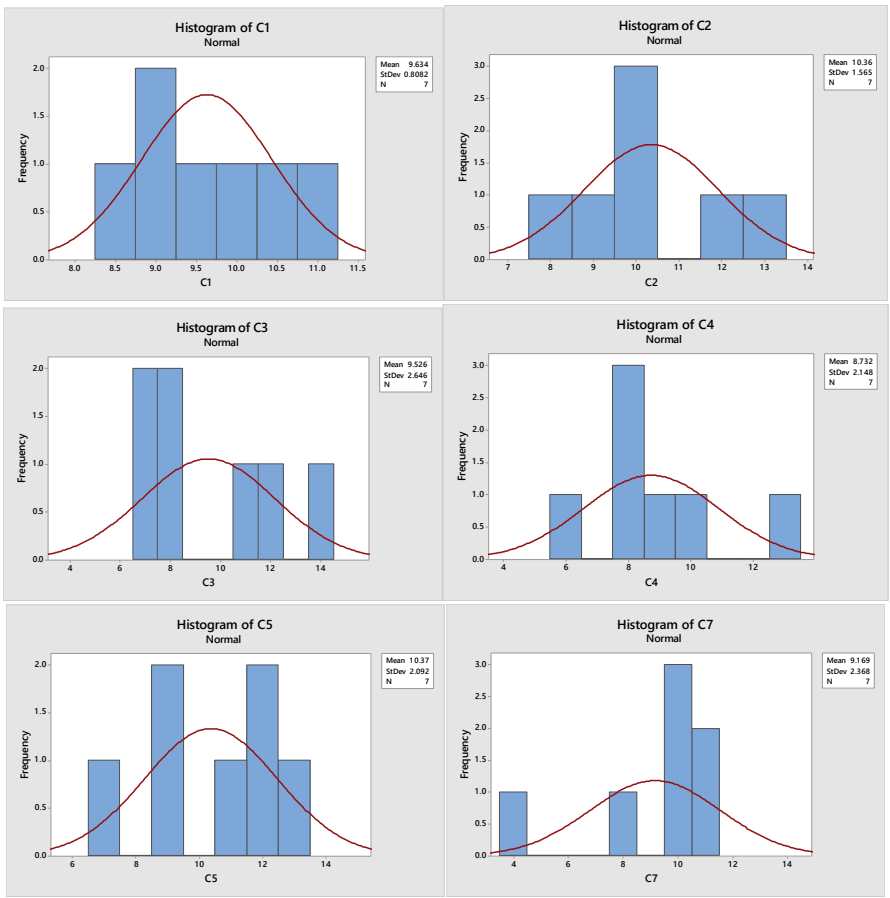
[22] This is unlikely to happen in a purely science research, but it not beyond imagining for (say) state or federal government folks, who often need to keep track of many variables, and may only be able afford a small effort on each.

[23] 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.

[24] 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 6 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[25]?

Figure 6. 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.

[25] 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)[26] 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 = two[27] (never mind that n = 2 sucks). We turn now to an introductory treatment of bootstrapping.

Section Four: Bootstrapping with two independent samples.

Bootstrapping is useful when (1) the statistic you are using doesn't fit into the classical canon of Normality-based methods[28], 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[29]. 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_1 sample values an infinite number of times[30]. Repeat for the other sample (of size n_2). This will generate two populations (of infinite size) from which to do the simulation.
(2) Draw a random sample of size n_1 from the first population, and n_2 from the second. Write down the difference in means.

---

[26] 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.
[27] Strictly speaking, one can do so for n = 2 also, but the resulting intervals are pretty much useless, despite being validly constructed.
[28] An example might be the difference or ratio of two medians.
[29] 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).
[30] You might immediately 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,[31] although our Excel tools use B = 10,000, just because we can without inducing long waiting times[32]).
(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[33]. 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[34]? Read on.

Practical Algorithm

Replace steps (1) and (2) above with: draw random samples of size n_1 from the first sample, and n_2 from the second one, *sampling with replacement*. The next two steps are the same.

What does "sampling with replacement" mean, and why does it work? First, the "what," illustrated with a simple example. 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 in our paired data chapter). 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[35] 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 sparrow size 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

[31] 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.
[32] 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.
[33] 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.
[34] We will have to face it eventually, but hopefully only after a reasonably long and good life.
[35] Ken *did* get precisely this using a random number generator.

classical ones. This provides evidence that the distribution of the means is not particularly Normal, suggesting that 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 (*t*-based)**
80%	(3.00, 17.81)	(2.46, 17.71)
95%	(-0.89, 20.89)	(-1.73, 21.9)
99%	(-3.76, 26.54)	(-5.71, 25.88)

Appendix: Two equal or not two equal (variances, that is)

This appendix might be of immense interest to a few, and about zippo to most others. Your reading it comes with the confession that you are both a statistics aficionado, and, possibly, a nerd[36].

The authors discussed here were selected by Ken, based on

(1) he has their book on his bookshelves, and
(2) they are individuals whom he has respects for their work as statisticians, and for the contemporary ones, their work as teachers of statistics.

So… nonrandom, convenience sample. Take it for what it is worth.

Let's begin with our forefathers.

Ronald Aylmer Fisher.

Ronald Fisher worked as an agricultural geneticist, working at Rothamsted Experimental Station in the UK in the 1920s. The inventor of the F test (used in ANOVA and regression modeling) among other things, he is recognized[37] as "a genius who almost singlehandedly created the foundations of modern statistical science" and "the single most important figure in 20th century statistics"[38] The foregoing was cribbed from the Wikipedia entry on R.A[39].

In his book, "Statistical Methods for Research Workers" (seventh edition, 1938), Fisher espoused that one should assert and use equal variances for testing the difference between two means. Indeed, he makes clear (page 130) that this is not merely an assumption:

> "It has been repeatedly stated, perhaps through a misreading of the last paragraph, that our method involves the "assumption" that the two variances are equal. This is an incorrect form of statement; the equality of the variances is a necessary part of the hypothesis being tested, namely that the two samples are drawn from the same Normal[40] population."

[36] We authors cheerfully confess to both.
[37] Anders Hald. 1998. A History of Mathematical Statistics From 1750 to 1930. Wiley, New York. 795 pages.
[38] Bradley Efron. 1998. R.A. Fisher in the 21st Century. Statistical Science 13(2): 95-122. We note that Brad (inventor of the bootstrap) might also be regarded as one of the most important figures in 20th century statistics.
[39] https://en.wikipedia.org/wiki/Ronald_Fisher
[40] Ken added the capitalization to the word "normal" to highlight it as the name of a specific distribution, rather than a synonym of, for instance, the word "usual". We note also that currently, we would not be inclined to insert the word "Normal" in there at all, as asserting that the data come from a Normal distribution is not a requirement: the Central Limit Theorem, given a sufficiently large sample size, assures us that means, and differences in means have (at least approximately) a Normal distribution.

Fisher did not address the equal variance question in the context of confidence intervals (i.e. estimating the size of a difference). Back in the day, the focus was almost exclusively on testing.

George Snedecor and Bill Cochran

George Snedecor made many contributions to the foundations of statistical methodology; he founded the first academic department of statistics in the United States, at Iowa State University. His 1938 text **Statistical Methods** became an essential resource. The foregoing was cribbed from the Wikipedia[41] entry on George.

In the fourth edition of his book (1948), George repeated Fisher's argument in favor of asserting equal variances for testing the difference in means from two independent samples (page 82). Like Fisher, he did not address estimation. William Cochran, also a statistician at Iowa State University, joined with George for the sixth edition of the text; David F. Cox (also at Iowa) drove the work on the seventh and eighth editions, both being done after the passing of George and Bill.

In the eighth edition, they initially stick with the equal variances approach (pages 89 – 91), but then they later include instructions for handling the case of unequal variances, with the caveat that one ought to test for equality; if rejected, used the more complicated associated arithmetic (pages 96 – 98).

Contemporary authors

Here we skip the biographical details, and simply report their advice.

Sokal and Rohlf (1995)[42] was written back when folks had to do computations by hand, so a considerable portion of the text is devoted to instructions on calculations (not so necessary these days, we think). That said, they show both how to do a test assuming equal variances (Calculations Box 9.6, page 225) and without (Box 13.4, pages 404-405). We stand with Fisher and Snedecor and would do testing with the *assertion* of equal variances. They show C.I. construction for the case of equal variances (Box 9.6); they do not address intervals for the unequal variance case.

Zar (2010)[43], also heavy on computational details, uses the equal variance approach for testing, but discusses unequal variance case as a violation of assumptions, and gives suggestions for dealing with it (page 136). Notice that dealing with equal variances as an assumption is different from simply asserting it (à la Fisher, Snedecor, and yours truly). He then goes on to give detailed instructions for choosing between the two (pages 141-

[41] https://en.wikipedia.org/wiki/George_W._Snedecor
[42] Robert Sokal and James Rohlf. 1995. **Biometry** (3rd Edition). Freeman and Company, New York. 887 pages.
[43] Jerrold Zar. 2010. **Biostatistical Analysis** (5th Edition). Pearson Prentice Hall, New Jersey. 994 pages.

142). For estimation via confidence intervals, he shows how to do it both ways (pages 142-145).

Ramsey and Schafer (2013)[44] use the pooled SE (page 40) so that the discussion keeps the "equal variance" flavor that will come along in regression and ANOVA. An arbitrary argument, but that's OK. Indeed, the two-independent sample case for estimating differences in means has a direct analogue when estimating proportions. For the two-sample case for estimating in proportions, it is common practice these days to argue that one should pool the data to estimate SE for the test, but use independent estimates of SD for the estimation (i.e. confidence interval) side of the business. Indeed, these authors do precisely that (page 556). So they could have just as well argued for pooled SE for testing the difference in means, and unequal variances for estimating the difference.

De Veaux, Velleman, and Bock (2013)[45] argue that assuming equal variances adds an unnecessary assumption, and so use the "don't assume…" approach for both testing and estimation. (pages 552 – 562). They show testing with the equal variances approach, but insist they are doing so only out of historical interest.

Utts and Heckard (2012)[46] show both in detail, but then caution that the pooled approach rests on a critical assumption and so should be used cautiously. (p. 518)

So what to make of all this?

- **(1)** You can do whatever you want: insist on always using equal variances, or always not, or using them differentially. No matta. You can find an expert (see above) ready to back you up.
- **(2)** We still like
  - **a.** Equal variances for testing because doing so is consistent with the null hypothesis
  - **b.** Don's assume equal variances for the estimation part. It amounts to an unnecessary assumption.

A note… For estimation, if the variances are clearly *not* equal, then the "assume equal variances" approach is wrong, simply. If they are reasonably close in value, and you use the equal variances approach and we use the other, the resulting confidence intervals will be almost identical. **So:** the equal variance approach is either wrong or gets you the same answer anyway. Why use it? End rant.

Another note: The degrees of freedom formula under the condition of equal variances is total sample size minus 2. Without that assumption, a more complicated formula[47], is used. It will produce a smaller number. To the degree the variances differ, to that degree Satterthwaite's degrees of freedom are smaller than that obtained using the simpler formula.

---

[44] Fred Ramsey and Dan Schafer. 2013. **The Statistical Sleuth** (3rd Edition). Brooks/Cole, New York. 760 pages.
[45] Richard De Veaux, Paul Velleman, and David Bock. 2013. **Intro Stats** (4th Edition). Pearson, New York. 815 pages.
[46] Jessica Utts and Robert Heckard. 2012. **Mind on Statistics** (4th Edition). Cengage, New York. 717 pages.
[47] Due to David Satterwthwaite.