Mastering Confidence Intervals

This module is organized into three “levels” of increasingly deeper understanding and sophistication of use of confidence intervals. Each builds on the previous. Trying to tackle Level Three before mastering Levels One and Two may prove frustrating. Full understanding of confidence intervals requires mastery of “sampling distributions.” I will touch on that topic here but recommend working with that module alongside this one.

Level One is designed to help you gain competence principally as a consumer (reader, user) of confidence intervals. What is the correct interpretation of a confidence interval? What makes confidence intervals wider or narrower?

In Level Two you will gain competence in basic construction. In many instances, statistical packages will do computations for you; you need to know which tools to use, when their use is valid, and how to interpret the results. For deeper understanding of the general principles, and to enable you to calculate them yourself when necessary, Part A will instruct you in confidence interval calculations in a few common settings. The principles (but not all the details) will carry over to a much wider set of circumstances. Part B will help you to understand why the calculations work.

Mastery of Levels One and Two will suffice for routine statistical work wherein interest is focused on means, proportions and differences between means and between proportions. With modern computing capabilities, you can ask a much broader array of questions for which there were previously no readily available tools. Level Three will introduce you to C.I. methods for such questions, in particular the use of the bootstrap, a data-driven simulation method. Use of the bootstrap is becoming common, so some familiarity with it will stand you in good stead. I will limit discussion to an introduction (with examples) of the basic ideas. Going beyond that deserves its own module (in preparation).
Level One: Consumer

Upon completion of this section, we hope you will

- Understand the properties of confidence intervals
- Know how to interpret a confidence interval
- Understand the concept of “margin of error”
- Understand the role of sample size and confidence level
- Amaze and impress your family with how much you know about statistics

Mastering these elements will enable you to confidently use confidence intervals as a reader of scientific articles, output from a statistical package, and other media (they appear more and more commonly in newspapers, for instance).

Example 1.1. From USA TODAY June 27, 2006. Only 40% of consumers plan to take a vacation in the next 6 months, a 28-year low. The estimate has a margin of error of plus or minus 3 percentage points, accurate 19 times out of 20. What does that statement mean to you?

Example 1.2. In a study\(^1\) to see the effects of calorie reduction on longevity, mice in one group were fed a post-weaning diet of a standard 85kcal/wk. Mice in the calorie-reduced group were, post-weaning, held at 50kcal/wk. On average, the calorie-reduced mice lived 9.6 months longer than the control group. We are 95% confident the calorie reduction increases lifespan somewhere between 7.3 to 11.9 months. What does that statement mean to you?

We could at this point simply list the relevant features and behavior of confidence intervals, but it will be more meaningful if you first see them in action.

---

Properties of a Confidence Interval

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. 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. Here we will use 90% as my choice, both to be rebel, 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 1 are the confidence intervals from twenty such samples. Notice that 18 of the 20 actually include the true mean; 2 of them miss it.

This illustrates the technical defining property of confidence intervals: if you repeatedly use the procedure, 90% (my choice here) of the resulting intervals will in fact contain the parameter being estimate. In other words, 90% of them “work”. 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.

---

2 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.

3 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%.
Figure 1. 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.

Biologist number\(^4\) 10 (labeled as such in the foregoing graph) is likely pretty pleased with her confidence interval; after all, it is quite narrow\(^5\), but notice that the interval does not contain the truth. Biologist number 1, on the other hand, may be a tad dismayed at his rather wide interval, but notice that his sample mean is almost dead on the population mean\(^6\).

In the notes that follow, I will use 95% as the named confidence level, since it is the one most commonly used in science, but the statements apply more generally.

1. A 95% confidence interval, if properly constructed, has the property that in a long run of using such intervals, 95% of them will contain the population value being estimated. We typically don’t know if any one interval includes or excludes that value but can only hold a certain confidence that it does.

2. In a classically constructed confidence interval, the lower limit equals the estimate minus the margin of error, and the upper limit equals the estimate plus the margin of error. Short-hand for this: estimate \(\pm \text{MoE}\)

---

\(^4\) Us statisticians go through biologists so quickly, we usually just assign them numbers instead of names 😁.

\(^5\) The fact of it being narrow is that, by chance the SD in her sample was quite small.

\(^6\) He, of course, will never know, since all he has is his sample (obviously, if we knew that the population mean was 50, we wouldn’t be sampling to estimate it).
Example 1.1 continued. Newspaper reports of polls don’t usually report the interval, just the estimate and the MoE. Here, a 95% lower limit is 40% - 3% = 37%; the upper limit is 43%.

Example 1.2 continued. In science reports, the C.I. is often given directly. For this one, we can deduce that the margin of error is 2.3 months.

(3) The margin of error is determined by
• choice of confidence level (lower will make the MoE smaller),
• choice of sample size (an increase will make the MoE smaller), and
• innate variation in the data (less will make the MoE smaller).

Here are more details on these behaviors. We summarize them in the following notes, as illustrated in Figure 3.

(1) A 90% interval will be narrower than a 99% 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 more sure 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 $n = 30$).
Two of these three the researcher can in principle control. I say, “in principle,” because “95%” is so culturally ingrained that using a different level usually requires some rationale, whereas if you pick 95%, no one will question your choice. Sample size is often not freely chosen because of logistical, budgetary or other constraints. It is rare that you can do anything about the amount of variation in your data.
Interpretation of a confidence interval.

The following are common interpretations of a confidence interval among folks just learning about them, not all of them correct. Consider each of the following statements about a standard 95% interval for a population mean. If you are sure you know which are true and which false, read on. If you are unsure, take a look at the CI quiz Excel file (in the Excel Tools folder) and take the interactive quiz.

There are two possibilities for taking the next steps deeper into confidence intervals. One is to learn how to compute them (at least, in statistically simple settings); the other is to learn more deeply the underpinnings for our claims about their properties. Level Two explicitly considers these two aspects.
Level Two: Standard Intervals

For a wide array of commonly used statistical methods, confidence interval construction follows a very straight-forward pattern with differences principally in the details. Any time we make a statistical estimate from some data we do so with a statistic (some number calculated from our data); our data ideally are values in a random sample from some population of interest. We use the statistic to estimate a population parameter (some number that would be calculated from the population if it were possible to observe all the values). The chosen statistic is called an estimator; and an observed value of it an estimate.

Classical confidence intervals for a population parameter\(^7\) take this form: estimate minus and estimate plus margin of error. The margin of error is determined by several factors, including the choice of confidence level, sample size, and innate variability in the data. At the heart of the margin of error is a measure of variation of the estimator called the standard error of the estimate. I’ll give a detailed discussion of standard error in the sampling distribution module (this might be one of those points where you need to return after reading that module).

In order to keep the exposure of these second-level details in doses small enough that your immune system doesn’t get rattled, I’ve written the Level Two material in two subsections. Part A is devoted to how to construct standard intervals; Part B is devoted to why they work the way they do. Further, I will gloss over procedure details for the sake of keeping attention on the big picture. In the fullness of time (i.e. if you take the time to study beyond these initial modules), those details will become better known to you. Part A of Level 2 follows immediately on the next page. Part B follows that.

\(^7\) Notice that we make an interval for a parameter, using a statistic. So when folks say, “a 95% confidence interval for the mean is….,” they are speaking of the population mean, not their sample mean. After all, they know what their sample mean is.
Level Two (A): Basic Construction

Upon completion of this section we hope you will know how to calculate confidence intervals for simple situations.

We will go through the details for four examples. Together they will illustrate the principles that don't change and the details that do when forming confidence intervals. I'll write up each one in similar detail; pick one to read first and slowly. By the time you read the fourth, I hope you are reading more quickly and seeing the patterns among the interval constructions (and where they differ).

**Question for Example 2.1** (Based on Example 1.1). In a random sample of 1000 consumers, 40% claimed that they would NOT be taking vacations in the next six months. What is a 95% confidence interval for this proportion?

**Question for Example 2.2** (derived from Example 1.1). In a random sample of 1000 consumers, 40% claimed that they would NOT be taking vacations in the next six months. In a similar random sample of 1000 (different) consumers taken a year earlier, 48% made that statement. What is a 90% confidence interval for the difference in those proportions?

**Question for Example 2.3** (from same study as Example 1.2). The lifespan for the 57 mice on the standardized diet was 32.7 months; the sample SD was 5.125 months. What is a 95% confidence interval for the true average lifespan of mice like these?

**Question for Example 2.4** (from same study as Example 1.2). The lifespan for the 57 mice on the standardized diet was 32.7 months; the sample SD was 5.125 months. The lifespan for the 71 mice on the reduced calorie diet was 42.3 months; the sample SD was 7.77 months. Calculate a 95% confidence interval for the true difference in mean lifespans of mice on diets like these (this is the interval reported in Example 1.2).

---

8 Reality Check: Polls like this, properly done, are not done with simple random samples. To do so would require a list of all eligible consumers in the U.S., which is not feasible. Appropriate (and likely used) sampling methods are more complicated. That means that at some point the confidence interval calculations are also more complicated than the ones I'll use here. I'll let you know where that occurs.
Confidence Interval Calculation Examples

Example 2.1 We have a presumed simple random sample, \( n = 1000 \). The sample proportion is \( \hat{p} = 400/1000 = 0.40 \) (or 40%). Goal: a 95% C.I. for the true proportion.

Our efforts here are aimed at estimating the proportion \( p \) (of consumers not vacationing in the next six months) in the entire population from which the sample is drawn. The sample proportion \( \left( \hat{p} \right)^9 \) is used to estimate that parameter. Each datum here is a “yes” or “no” (recorded as a 1 or a 0), so \( \hat{p} \) is a special kind of proportion called a Binomial proportion.

The standard error of a Binomial proportion from a simple random sample is estimated by\(^{10} \)

\[
SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

The MoE for a standard interval\(^ {11} \) for a proportion is computed as \( z_{cl} \times SE(\hat{p}) \), where \( z_{cl} \) specifies how many SEs wide the MoE needs to be. It comes from a standard Normal distribution (for details, see Part B). The value sought is \( z \) such that \( CL\% \) of the distribution is captured between \( z \) and \( -z \).

For a 95% confidence interval, \( z_{.95} = 1.96 \) and the SE is

\[
SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.4 \times 0.6}{1000}} = 0.015
\]

Hence the MoE is \( 1.96 \times 0.015 = 0.03 \). We are 95% sure the true proportion is between \( 0.40 - 0.03 = 0.37 \) (37%) and \( 0.40 + 0.03 = 0.43 \) (43%).

Example 2.2 We have two simple random samples taken a year apart, each with \( n = 1000 \); \( \hat{p}_1 = 48\% \) in the first sample, and \( \hat{p}_2 = 40\% \) in the second. Goal: a 90% C.I. for the true difference.

Our efforts here are aimed at estimating the difference in proportions \( \left( p_1 - p_2 \right) \) (of consumers claiming they had no vacation plans in the subsequent six months) in the population of eligible consumers. The statistic chosen to estimate the parameter is quite naturally\(^ {12} \) \( \hat{p}_1 - \hat{p}_2 \).

The standard error is estimated by

\[
SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}
\]

\(^9\) For some parameters, the notational convention is to choose a letter to denote the parameter; the same letter with a carat (\(^\wedge\)) over it denotes the estimator.

\(^{10}\) The SE formula for a proportion from a more complicated random selection scheme is itself more complicated.

\(^{11}\) Reality check: In recent years, there has been a batch of interesting papers in the statistical literature demonstrating the poor performance of the standard method. I’ll save that discussion for the module on Binomial proportions. I’m using it here because (1) it is easy to understand, and (2) you will see it in use in textbooks and research papers for several more years. We scientists are slow to change our ways.

\(^{12}\) The order of subtraction is not important; only that, once it is chosen, take care to follow through consistently. I chose to subtract the second from the first in order to get a positive number for an answer (minus signs confuse me 😞).
standard C.I. for the difference in proportions is computed as \( z_{CL} \times SE(\hat{p}_1 - \hat{p}_2) \), where \( z_{CL} \) specifies how many SEs wide the margin of error needs to be. It comes from a standard Normal distribution (for details, see Part B). The value you seek is \( z \) such that \( CL\% \) of the distribution is captured between \( z \) and \(-z\).

For a 90\% confidence interval, \( z_{90} = 1.645 \) (verify this). The SE for our example is \( SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{.48(.52)}{1000} + \frac{.4(.6)}{1000}} = 0.022 \). The MoE is \( 1.645 \times 0.022 = 0.036 \). We are 90\% sure the true difference in proportions is between \( 0.08 - 0.036 = 0.044 \) (4.4\%) and \( 0.08 + 0.036 = 0.116 \) (11.6\%).

**Example 2.3** We have: a presumed simple random sample, \( n = 57 \); sample mean is \( \bar{y} = 32.7 \) months; sample SD\(^{13}\) is 5.125. Goal: a 95\% C.I. for the true mean.

Our efforts here are aimed at estimating the mean \( \mu \) in the entire population from which the sample is drawn. The sample mean \( (\bar{y})^{14} \) is used to estimate that parameter. The standard error of a mean from a simple random sample is estimated by \( SE(\bar{y}) = \frac{SD}{\sqrt{n}} \).

The MoE for a standard C.I. for the mean is computed as \( t_{n-1,CL} \times SE(\bar{y}) \), where \( t_{n-1,CL} \) specifies how many SEs wide the margin of error needs to be. It comes from a \( t \) distribution with \( n - 1 \) degrees of freedom (noted in the subscript) (for details, see Part B). The value you seek is \( t \) such that \( CL\% \) of the distribution is captured between \( t \) and \(-t\).

For a 95\% confidence interval, \( t_{56,0.95} = 2.00 \) (verify this with the table) and the standard error is \( SE(\bar{y}) = \frac{5.125}{\sqrt{57}} = 0.68 \). Therefore the MoE is \( 2 \times 0.68 = 1.36 \). Hence we are 95\% sure the true mean lifespan is between \( 32.7 - 1.36 = 31.34 \) months and \( 32.7 + 1.36 = 34.06 \) months.

\(^{13}\) SD is a conventional shorthand for standard deviation. Sometimes simply the letter \( s \) is used for the sample SD; the population SD it estimates is typically denoted by \( \sigma \).

\(^{14}\) For some parameters, the convention is to denote the parameter by a Greek letter (here, \( \mu \) because it coincides with the “m” in “mean”). Some letter choice is used to denote a variable ( \( y \) and \( x \) are common), and a bar over the letter denotes the simple mean of the values in the sample. Individual values in a sample are accounted for by subscripting: \( y_1, y_2 \), and so on with \( y_i \) (other letters are used, but often \( i \) ) to denote an unspecified (or generic) value.
Example 2.4 We have: (presumed) simple random samples, \( n_1 = 57, n_2 = 71 \); means are \( \bar{y}_1 = 32.7, \bar{y}_2 = 42.3 \); sample SDs are \( s_1 = 5.125, s_2 = 7.77 \). Goal: a 95% C.I. for the difference in the true means.

The parameter of interest is the number \( \mu_2 - \mu_1 \), the difference in population mean lifespans between the standard diet (group 1) and reduced calorie diet (group 2). The sample difference\(^\text{15} \) \( (\bar{y}_2 - \bar{y}_1) = 9.6 \) is used to estimate that parameter.

The SE of the difference in two independent means from simple random samples is estimated by\(^\text{16} \) \( SE(\bar{y}_2 - \bar{y}_1) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \).

The MoE for a standard interval for the mean is computed as \( t_{df, CL} \times SE(\bar{y}_2 - \bar{y}_1) \), where \( t_{df, CL} \) specifies how many SEs wide the margin of error needs to be. It comes from a \( t \) distribution with degrees of freedom that come from a complicated formula, which is why we use computer packages. For now, I’ll just tell you that \( df = 121 \) for this data set. With \( t_{121,0.95} = 1.98 \) (verify this) and \( SE(\bar{y}_2 - \bar{y}_1) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{5.125^2}{57} + \frac{7.77^2}{71}} = 1.145 \), the margin of error is \( 1.98 \times 1.145 = 2.27 \). Hence we are 95% sure the true difference in mean lifespans is between \( 9.6 - 2.27 = 7.33 \) months and \( 9.6 + 2.27 = 11.87 \) months.

\(^{15}\) The order of subtraction is not important, but once it is chosen, take care to follow through consistently. I chose to subtract the first from the second to get a positive number for an answer (simpler interpretation).

\(^{16}\) Notice I’ve changed notation for standard deviation from SD to s. The two are often used interchangeably…
Level Two (B): Underpinnings

Upon completion of this section I hope you will

- Understand the role of sampling distributions in motivating confidence intervals
- Understand more deeply the role of sample size, confidence level and random variation on the margin of error

The Sampling Distribution of a Statistic

For the sake of being concrete, I will use here a confidence interval for a mean. Where an explicit confidence level is helpful, I will (boringly) use $CL = 95\%$.

The formula for a 95% confidence interval for the mean is

$$\bar{y} \pm t_{n-1,CL} \times SE(\bar{y})$$

(see Example 2.3 for a detailed example with numbers). The logic behind this formula hinges on the imaginary act of repeating our study ad infinitum. In that imaginary act, we would generate all possible sample means that could arise from studies exactly like ours. We imagine, then, that there is some distribution of those means. This (usually imagined, rarely seen) distribution of a statistic is called its sampling distribution.

It is often reasonable to assume that the sampling distribution in question is approximately Normal (the $t$-distribution is functionally a tweak on the Normal distribution$^{17}$). In many instances for biological data, the population of values that gave birth to the sample is emphatically not Normal; why then can we claim that the sampling distribution of (say) the mean is approximately Normal? I’ll save a more nuanced discussion of that question for the module on sampling distributions; for our purposes here, I’ll demonstrate that the sampling distribution of a sample mean is often approximately Normal property, and then give it a name.

Simply put, (1) the distribution of a mean from a random sample gets more and more like a Normal distribution as sample size increases, and (2) it gets there more easily (i.e. for smaller sample size) when the original distribution is itself more symmetric. In this note, I focus on the confidence interval for a mean, but it might be interesting to take a look at a similar phenomenon for the sampling distribution of a Binomial proportion. In that case, (1) the distribution of a binomial proportion from a random sample gets more and more like a Normal distribution as sample size increases, and (2) it gets there more easily (i.e. for smaller sample size) when the underlying proportions are closer to 50% (i.e. the original distribution is itself more symmetric).

When basic statistical tools were being developed, this property of the sampling distribution of many statistics approaching Normality with increasing sample size caught the attention of statisticians. Formal study of it led to what we now call the Central Limit Theorem, which, at its simplest$^{18}$, asserts (and proves) the result. The CLT, as it is known in short-hand, forms the theoretical justification for much of routine statistical inference.

---

$^{17}$ In particular, it adjusts the shape of the distribution to account for the fact that the underlying SD is imperfectly estimated.

$^{18}$ As simple cases got pinned down, and the result proven, statisticians worked at expanding the conditions under which it (or some variation of it) works. The result is that there are actually many many many versions of this theorem.
In days gone by (before computers were around to study such things easily), sample size 30 was often invoked as being about the smallest sample size that would give the user comfort that the CLT was working (i.e. that the \( t \) -test he or she was about to use was valid). A single, universal value is not often espoused any more; we know, from simulation studies (and theoretical research) that there are circumstances in which quite less is sufficient, and others when 30 is nowhere near enough. I will defer detailed discussion of that question for relevant modules on particular statistical tools.

**Sample Size, Confidence Level, and Variation**

Most of what I wish to convey here you studied in an experiential way in Level One; here, I will summarize the main points to tie it all together. I will continue with the concrete case of a confidence interval for the mean.

The first term in the margin of error \( \left( t_{df,CL} \right) \) essentially answers the question, “How many standard deviations (of the sampling distribution of the mean) wide does the interval need to be?” The \( df \) in the subscript reminds us that the correct \( t \)-distribution to use is the one with \( df \) degrees of freedom. The other part connects your choice \( CL \) of confidence level to that \( t \)-distribution. For example, the correct value for a 90% interval for the mean from a random sample of 57 values (as per Example 2.3) is \( t_{56,90\%} = 1.67 \), whereas for a 95% interval with those same data, it is \( t_{56,95\%} = 2.00 \).

The number you use to calculate the confidence interval will depend quite strongly on the confidence level and less (for moderate to large sample sizes) on the sample size. The other element in the margin of error is the so-called standard error of the statistic, which for the mean from a random sample is \( \frac{SD}{\sqrt{n}} \) (\( SD \) is sample standard deviation, and \( n \) is the sample size).

Other statistics will have other formulae, but the pattern that remains constant is comprised of (1) the (relevant form of the) variation in the data will be in the numerator, and (2) the sample size (in one form or another) will be in the denominator. This ought to be intuitively appealing: the more variable are the data, the less precise our estimates; the larger our samples, the more precise our estimates.

Confidence intervals for many commonly used statistics take the form we’ve been discussing so far in Levels One and Two. We can ask a broad array of questions these days for which formal statistical inference (i.e. making intervals or doing tests) was not possible (or, at least, was far from easy) in days gone by. Thanks to modern computer technology and recent advances in computer simulation-based statistical tools, we can relatively easily do inference for many of those questions. In Part Three, I will introduce you to a widely applicable tool for answering those questions and attempt to convince you that doing inference for means, and differences in means, is unnecessarily restrictive scientifically.

---

\(^{19}\) Recall that the SE of a statistic is an estimate of the SD of the sampling distribution of that statistic.
Level Three: Confidence Intervals for Nonstandard Situations (Brief Introduction to the Bootstrap)

Upon completion of this section I hope you will
- Understand the conceptual underpinnings of the bootstrap
- Know how to interpret results of bootstrap procedures

Nonstandard Situations

Under what circumstances would one of the standard statistical tools, routinely available in statistics packages, not be adequate? In general terms, if you are dealing with a situation where you believe the sampling distribution of the statistic may be non-Normal and/or there is no readily available method of estimating the standard error of the statistic. If no such situations come readily to mind, one reason is that the methods folks routinely use (based on means, Binomial proportions, and relevant differences between them) are pretty much all the ones that practitioners get exposed to in their statistics classes. In years gone by, that was defensible, since, after all, teachers should only teach methods that they are confident work well\(^20\).

I’m going to use the mouse longevity study examples (1.2, 2.1, and 2.3) to illustrate that examining differences in means may not be the most meaningful question to ask. Recall there were two diet groups: (1) a standard (85 kcal/week) diet and (2) a reduced-caloried (50 kcal/week) diet. The standard question (motivated by the most easily available tool\(^21\)) is whether there is a difference in mean longevity \((\mu_2 - \mu_1)\) between the two groups. The answer is an estimated gain in lifespan of 9.6 months, with a 95% confidence interval from 7.33 to 11.87 months. That’s fine, I suppose, but it’s not information that is immediately relevant to me\(^22\).

Lifespan data are usually skewed to the left\(^23\), as illustrated for the data from this experiment (a Normal curve is overlain for comparison).

\(^{20}\) History and inattention has played a small joke here on statisticians and practitioners alike, in that one of the very simplest tools in most statistical toolkits (and often one of the first methods taught in classes) is the standard confidence interval for a Binomial proportion: \(\hat{p} \pm z_{CL} \sqrt{\hat{p}(1 - \hat{p})/n}\). The joke is that it doesn’t work well at all (see module on proportions for details).

\(^{21}\) An example of the tail wagging the dog: after all, shouldn’t the scientist be free to pose the question of interest, and then seek an appropriate statistical tool?

\(^{22}\) For the record, I am not a mouse.

\(^{23}\) The distribution of many biological variables is skewed right; lifespan data are an interesting counter-example to that norm.
Skewed left implies that the means are less than the medians. Here are side-by-side boxplots (the horizontal line inside each box marks the median; the cross inside the circle, the mean).

The mere fact that the means and medians are different\(^2\) raises an interesting question: which population parameter (means or medians) is more interesting here? Let’s think about this problem in terms of human lives for a minute. Mean lifespan is connected to “total” in a way that might be useful for some purposes: mean times number of lives gives a sense of the total size of, say, the problem for a Social Security administrator. Median, on the other hand, is more representative of what you and I are likely to run into: for skewed populations; median represents “typical” in a way the mean does not\(^2\). I’m going to argue the median is more interesting, and for sake of continuing to write this chapter (and to choose a “nonstandard question”), will take it as my parameter of interest.

\(^2\) Let’s presume this shape is reflected in the original populations (reasonable given the sample sizes).

\(^2\) This would all be moot if the distributions were symmetric, for then the mean and median are the same number.
It also occurs to me that a conclusion (for means or medians) that relays a relative difference (35% longer, for instance) yields a more dramatic and compelling punchline that saying simply that they are different by some number of months. I contend that biology is often more interesting on a relative scale than on a simple “difference” scale, and this is one such case. So, there it is: I would like a confidence interval for the ratio of medians. In this instance, it is not at all clear what the sampling distribution would look like, nor is there a handy way to estimate the standard error (methods do exist; note I said, “handy”).

**Bootstrapping to Estimate Standard Errors and Sampling Distributions**

This discussion presumes that you understand the basic ideas of sampling distributions. If that phrase is not clearly meaningful to you, I recommend working with the sampling distributions material in the third section in the Core Concepts Chapter to increase your understanding.

“The sampling distribution” is the answer to the question, “What would the distribution of my statistic look like if I repeated my study a very large number of times?” Standard \( t \) and \( z \) methods for means and proportions are based on the following premise: If the sample size is large enough and/or the population the sample came from is reasonably symmetric (a Normal distribution is the ideal), then it is reasonable to presume the sampling distribution of the statistic in question is approximately Normal.

The bootstrap is a data-based simulation method to answer the question, “What is the (estimated) sampling distribution of the chosen statistic?”. At its simplest (i.e. in the simplest setting), it works like this (some details omitted; see the module on bootstrapping for more):

1. Use the sample itself as a representation of the population from whence it came.  
2. Draw random samples from that representation (using the same sample size as the original sample); for each of these so-called bootstrap re-sample, calculate the chosen statistic. Repeat a large number of times (say, 1000).  
3. The histogram of the resulting estimates of the statistic is a simulation-based estimate of the sampling distribution of the statistic.  
4. Then, for instance, sorted from smallest to largest, the 26\(^{th}\) (of 1000) ordered values is the lower limit of a 95% empirical bootstrap confidence interval (2.5% are smaller), and the 975\(^{th}\) is similarly the upper limit.

In our mouse longevity study, for each bootstrap replicate, we would draw a sample from the representation of each population and then calculate our chosen statistic (ratio of medians, say). In the end, we would have a 95\% () confidence interval for the ratio of medians (or whatever confidence level we desire). The following Excel tool does that calculation. Note that it

---

26 Look above at the histograms of the two samples for the mouse data. The population of lifespans for the standard diet looks (roughly) 3% are around 18 months, 6% are around 21 months, and so on.  
27 We call it an estimate even though it is not just a single number; the entire distribution has been estimated.
allows you to choose means or medians, differences or ratios and “order of operation” (i.e. do you want median 1 divided by median 2 or the other way around?). It also supports analyses for paired data (which this mouse study is not).

**Cut to the Chase:** We are 95% sure the reduced diet mice typically (using the medians) live from 30% to 47% longer (30% and 47% are the lower and upper C.I. limits the bootstrap procedure created). Now, *that* is an interesting conclusion!