Chapter 6.2 - Confidence Intervals

> “Uncertainty is a personal matter; it is not the uncertainty but your uncertainty.” - Dennis Lindley

The problem with using statistics in practice is that everyone wants *point estimates* (single value answers) when it can be argued that those don't even *exist*. If you were to go to an engineer at NASA and ask them what the value of π is they would say it's 3.141592653589793, because at that degree of accuracy you could launch an object across the observable universe and, at worst, miss the target by a couple inches. If you went to a mechanical engineer they would say $\pi = 3.14$ since anytime they're working without a calculator there's no need for accuracy, those results won't be final. This is because engineers almost exclusively exist in the world of point estimation where the *precision* of the estimate is the only measurement of its quality.

Statistics is a field that abandons pure point estimation. Rather than generate some hyper precise estimation of π, a statistician might provide an *interval estimate* of that same value. A good way of thinking about this is with rounding errors. Imagine you're asked to calculate $\pi/2$ but you're not told how precise the calculation needs to be. Instead of guessing the level of accuracy needed we could calculate an interval with a lower bound, a point estimate, and an upper bound:

$$\text{Lower Bound} = \pi/2 = 3.1/2 = 1.55$$

$$\text{Point Estimate} = \pi/2 = 3.14/2 = 1.57$$

$$\text{Upper Bound} = \pi/2 = 3.141592653589793/2 = 1.570796$$

We would then provide the answer of “$\pi/2$ is somewhere between 1.55 and 1.570796, but it's likely equal to 1.57”. While this is a crude explanation of the concept ahead of us, it's a useful analog to help us put on our interval estimation goggles and head into the storm that is uncertainty quantification.

> A physicist, a biologist and a statistician go hunting. They are hiding together in the bushes and they see a deer 70 ft ahead of them. The physicist makes some calculations, aims and fires at the deer. His shot ends up 5 ft to the left of the deer. The biologist analyzes the deer's movement, aims and fires. His shot ends up 5 ft to the right of the deer. The statistician drops his rifle and happily shouts, “We got it!”.

Pivots

Consider the sample of 10 deer below:

Location of harvest	Date of harvest	Sex	Age class	Body mass in kg
Desoto	2004-10-16	Female	2.5	45.8
Desoto	2004-12-12	Male	2.5	65.8
Desoto	2007-01-06	Female	4.5-5.5	44.5
Desoto	2005-12-11	Male	3.5	71.2
Desoto	2005-12-11	Female	4.5-5.5	42.2
Desoto	2005-01-09	Male	3.5	68.9
Desoto	2004-12-11	Male	2.5	61.7
Desoto	2010-01-02	Male	0.5	19.5
Desoto	2004-12-11	Male	4.5-5.5	70.8
Desoto	2007-01-06	Female	2.5	41.3

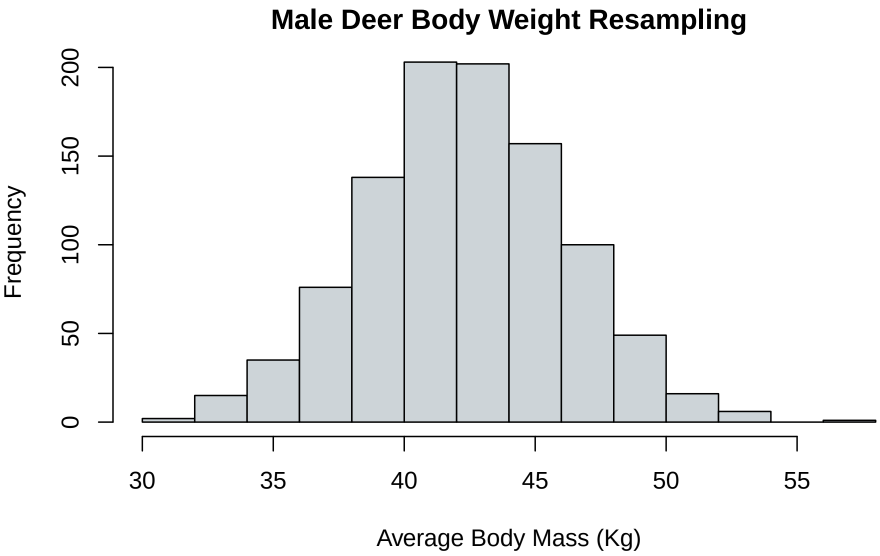
We know enough to be able to answer the question, “What is the average body mass of the deer in this sample?”.

$$\bar{x} = \frac{45.8 + 65.8 + 44.5 + 71.2 + 42.2 + 68.9 + 61.7 + 19.5 + 70.8 + 41.3}{10}$$

$$\bar{x} = \frac{531.7}{10}$$

$$\bar{x} = 53.17$$

With a little extra work (that won’t be shown) we could find the standard deviation of $s = 17.127$, among other useful sample statistics. We’ve learned how to fit a line through the data to describe the possible correlations between age or sex and body mass. If we wanted to talk about how much these estimates could possibly vary it might be helpful to use their *sampling distributions*.



What's nice about these sampling distributions is that our basic knowledge of distribution theory allows us to easily construct intervals for any specified probability.

$$P(-z_0 < Z < z_0) = \ 0.95$$

$$P(-1.96 < Z < 1.96) = \ 0.95$$

The problem is that we need to know the values of the population parameters for that sample statistic in order to describe its sampling distribution. Since population parameters are almost *never* observed in practice (hypothetically they're completely unobservable in all scenarios) we need to develop a way to construct these intervals using something that isn't *reliant* on the population parameters.

Our instinct might lead us towards the standard normal distribution. If a sample is **sufficiently large** then we should be able to standardize the values into z-scores and construct intervals that way. If we construct a statistic, z^* ("z star"), via the standardization of of x:

$$z^* = \frac{x - \mu}{\sigma/\sqrt{n}}$$

The statistic is still reliant on those pesky population parameters. But we can construct a similar statistic, t^*, by substituting in the sample statistics from the data for the population parameters:

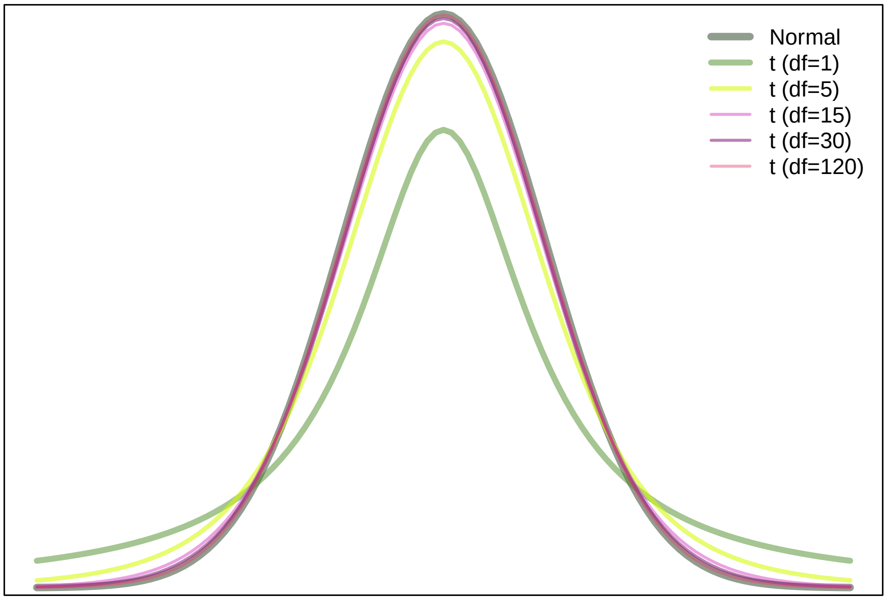
$$t^* = \frac{x - \bar{x}}{s/\sqrt{n}}$$

This statistic doesn't include the population parameters which makes it something called a **pivotal quantity** or **pivot** for short.

Pivotal Quantity (Pivot): Any function who's probability distribution does not rely of the population parameters of a given random variable.

Students of mathematical statistics spend quite a bit of time deriving pivots for random statistics and hypothetical data sets. While this is a wonderful exercise that anyone interested in a quantitative career should engage in, it's not necessary here. We're simply providing a little motivation to the otherwise bizarre steps we're about to take in constructing an interval estimate.

The key phrase in the definition of pivots is "probability distribution"; all pivots have a definable probability distribution. The notation of t^* was no accident either since the pivot we've constructed is t-distributed, that wonderful heavy-tailed curve Guinness blessed the world with.



The *only* parameter of the t-distribution is **degrees of freedom** (df) which has a very convenient formula for our situation:

$$df = n - 1$$

With this we have all we need to describe the distribution of our pivot and pull an interval estimate out of it!

Confidence Intervals

It'd be nice if we could create a general method for working with this new statistics, t^*, not because algebraic exercises are any amount of entertaining but because our new lives as statisticians has made *impeccably* lazy. So let's do that! We'll start by defining a "population" version of t^*, T:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

We'll substitute T for Z and t_0 for z_0 in our original interval estimation problem from chapter 4.1:

$$P(-t_0 < T < t_0) = \ 0.95$$

Substituting in the definition of T and solving for μ:

$$\begin{aligned} P(-t_0 < T < t_0) =& \ 0.95 \\ P(-t_0 < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_0) =& \ 0.95 \\ P(-t_0 \ \frac{s}{\sqrt{n}} < \bar{x}-\mu < t_0 \ \frac{s}{\sqrt{n}}) =& \ 0.95 \\ P(\bar{x} - t_0 \ \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_0 \ \frac{s}{\sqrt{n}}) =& \ 0.95 \end{aligned}$$

While we may have just brought back a population parameter after working **so hard** to get rid of them, this is all in the effort of generalization. What we have now is a simple formula for computing the interval for any t-distributed variable:

$$\bar{x} \pm t_0 \ \frac{s}{\sqrt{n}}$$

Recall how we developed the values of z_0 in chapter 4.1. We took the target probability, 0.95, subtracted 1 from it, and split it in half. If we consider a new probability, $P = 1 - \alpha$, we can split α in half to define a new value of t_0 that satisfies all potential intervals we might be interested in, $t_{\alpha/2}$. For example, let's define $P = 0.99$:

$$0.99 = 1 - \alpha$$

$$\alpha = 0.01$$

$$\alpha/2 = 0.005$$

We can use this trick for intervals from the standard normal distribution as well to let us find consistent values that satisfy probability intervals. With this we have our generalized formula for intervals of t-distributed variables:

$$\bar{x} \pm t_{\alpha/2} \ \frac{s}{\sqrt{n}}$$

This formula has two components to it— the *point estimate* and the *margin of error*.

$$\text{Point estimate} = \bar{x}$$

$$\text{Margin of error} = t_{\alpha/2} \ \frac{s}{\sqrt{n}}$$

Together this gives us the completed, generalized formula for **confidence intervals**, the frequentist method of quantifying uncertainty:

$$\text{Point estimate} \pm \text{Margin of error}$$

Intervals on Sample Statistics

When we worked with z-scores we had the z table. Since we're working with "t values", as we call them, we'll need to use the t table:

df	0.500	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.001
1	0.000	1.000	1.376	1.963	3.078	6.314	12.706	31.821	318.309
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	22.327
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	10.215
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	7.173
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	5.893
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	5.208
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	4.785
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	4.501
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	4.297
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	4.144

The rows of the table are separated by degrees of freedom and the columns are the probabilities we compute from the $\alpha/2$ method. Each of the cells is a "critical value" which is the value of $t_{\alpha/2}$ that satisfies the probability.

For example, in the deer sample we have $n = 10$ which means $df = 10 - 1 = 9$. If we want to construct a 95% confidence interval on the sample mean for deer body weights then $\alpha = 0.05$ and $\alpha/2 = 0.05/2 = 0.025$. Intersecting those two values in the table:

df	0.500	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.001
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	4.785
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	4.501
9	0.000	0.703	0.883	1.100	1.383	1.833	**2.262**	2.821	4.297
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	4.144

The critical value in this case is $t_{\alpha/2} = 2.262$. Plugging that into our formula for confidence intervals on t-distributed variables:

$$\text{CI} = \bar{x} \pm t_{\alpha/2}\ \frac{s}{\sqrt{n}}$$

$$\text{CI} = 53.17 \pm 2.262 \times \frac{17.127}{\sqrt{10}}$$

$$\text{CI} = 53.17 \pm 12.25107$$

$$\text{CI} = \begin{cases} 53.17 - 12.25107 = 40.91893 \\ 53.17 + 12.25107 = 65.42107 \end{cases}$$

$$\text{CI} = (40.92, 65.42)$$

We denote confidence intervals one of two ways, we can either hold the form of :

$$\text{Point estimate} \pm \text{Margin of error}$$

or use the numeric interval in the form of:

$$(\text{Lower bound},\ \text{Upper bound})$$

While both are valid it's typically best to stick with the latter for easier interpretation and usage of the interval in further calculations (something we'll encounter next section).

As discussed in chapter 6.1, the interpretation of this confidence interval would be: "As we take repeated samples from this population, calculate their sample means, and compute their confidence intervals, we would expect 95% of those intervals to cover the true population mean, μ". This formal interpretation can be troublesome for a lot of students and practitioners alike so statisticians tend to (falsely) apply the Bayesian credible interval interpretation of "There's a 95% chance that the true value is between 40.92 and 65.42". While the formal definition is quite un-approachable we shouldn't apply false definitions either.

The better way to interpret confidence intervals is to fuse the formal definition with the law of large numbers: "There's a 95% probability that further 95% intervals calculated will contain the true value of the population mean, μ". It's a tricky distinction but an important one. The interval we've created is only as valid as the method we've applied to the sample. It may seem odd to consider that the true value isn't inherently contained in the interval but that's the product of what we're quantifying the uncertainty for; the method itself.

Practically anything can have a confidence interval on it— although some are easier to calculate than others. In order to put a confidence interval on a proportion we have to make a couple changes to the formula, but the process remains the same.

Recall that proportions are generally derived from large samples and thus allow us to use the standard normal distribution instead of the t-distribution. Additionally the formula for variance of proportions is $p(1-p)$. We can substitute both of these into the confidence interval formula we derived previously and set $\hat{p}$ as the point estimate:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This is still assuming that the proportion is derived from a *reasonably* large sample, which we can assume as long as the CLT test is close or passing.

$$\hat{p} \sim N(\mu_{\hat{p}}, \sigma^2_{\hat{p}}) \quad \text{Given} \;\; np \geq 10 \text{ and } n(1-p) \geq 10$$

What's nice about confidence intervals on proportions is they're analytically simple. $z_{\alpha/2}$ doesn't change unlike $t_{\alpha/2}$, which changes depending on degrees of freedom, so we can calculate the common ones well in advance. The most common confidence intervals are 90%, 95%, and 99%. These intervals can be thought of as the level we're comfortable with a "mistake" occurring in our analysis, to an extent they're all "dealer's choice". If you're comfortable with mistakes popping up 1 out of 10 times then you would choose the 90% interval, 95% for 1 in 20, and 99% for 1 in 100.

As with most conventions in statistics, the popularity of 95% intervals is due to Fisher who chose to use it for his entire career because the value of $z_{\alpha/2}$ for 0.95 is close enough to 2 that he could round up and construct confidence intervals in his head.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	**0.9750**	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

This made 90% and 99% popular alternative since they were close to the 95% interval but could restrict or relax the standards by a lot. While other intervals are possible and equally valid we won't encounter them in the wild nearly as often. As such many statisticians tend to memorize the $z_{\alpha/2}$ values for those intervals:

$$z_{0.05} = 1.645$$

$$z_{0.025} = 1.96$$

$$z_{0.005} = 2.57$$

We can apply this formula for proportion confidence intervals to compute a 95% interval. Out of the 1797 deer in the data, 1105 of them are female which means the proportion of female deer is $\hat{p} = 0.615$. Despite having no knowledge of the true proportion, p, we can still construct a confidence interval to attempt inference about the population:

$$\begin{aligned}
\text{CI } &= \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
\text{CI } &= 0.615 \pm 1.96 \times \sqrt{\frac{0.615(1-0.615)}{1797}} \\
\text{CI } &= 0.615 \pm 0.0115 \\
\text{CI } &= \begin{cases} 0.615 - 0.0115 = 0.6035 \\ 0.615 + 0.0115 = 0.6265 \end{cases} \\
\text{CI } &= (0.6035, 0.6265)
\end{aligned}$$

This is where we can see the flaw in the logic behind confidence intervals. Deer tend to float around a 1 to 1 ratio with sex and this is a pretty big jump from that. This isn't the confidence interval or the sample being "wrong" it's our comprehension of it that's wrong. The 95% interval is a measurement of the method and the method was *computing a sample proportion.* The sample proportion considers only the data within the sample so we can reasonably say that if we repeatedly sample *roughly* the same number of deer from the population this technique would produce 95 out of 100 intervals containing the true value (which we could probably assume to be $\approx 50\%$).