Confidence intervals

We now want to take what we've learned about sampling distributions and standard errors and construct confidence intervals. What are confidence intervals? Simply an interval for which we have a certain confidence.

For example, we want to to be 90% certain that an interval that we construct, say, (a, b), contains the true value of something we're interested in (e.g. μ). A more specific example might be: we want to be 90% certain that the interval $(60, 75)$ contains the true mean height (μ) for men in inches.

Usually we are interested in confidence intervals for the mean, μ, but we can also construct confidence intervals for other parameters such as σ or p (the population proportion). However, we'll stick with confidence intervals for μ.

A few things we should notice. First, we calculate our confidence interval (CI) for μ, not $\bar{y}$. We *know* what $\bar{y}$ is, so there's no point in calculating a CI (since we know what it is, it doesn't have a CI in any case). We calculate a CI for μ precisely because we don't know what μ is, and we want to have some kind of limits for which we're reasonable sure about μ. Finally, remember that μ is not a random variable. It's a constant. That means we need to be very careful about how we define our limits. More on this last point below.

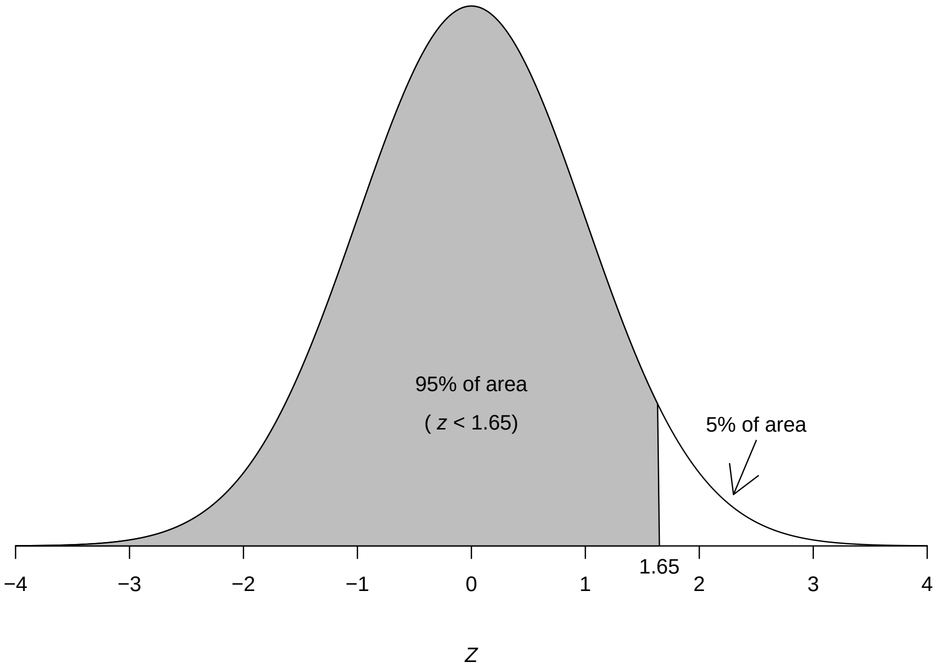
So how do we calculate a and b for our interval? Let's review some simple properties of the normal curve by looking at quantiles and figure out the probabilities associated with some common percentiles (quantiles) like 90%, 95%, and 99%:

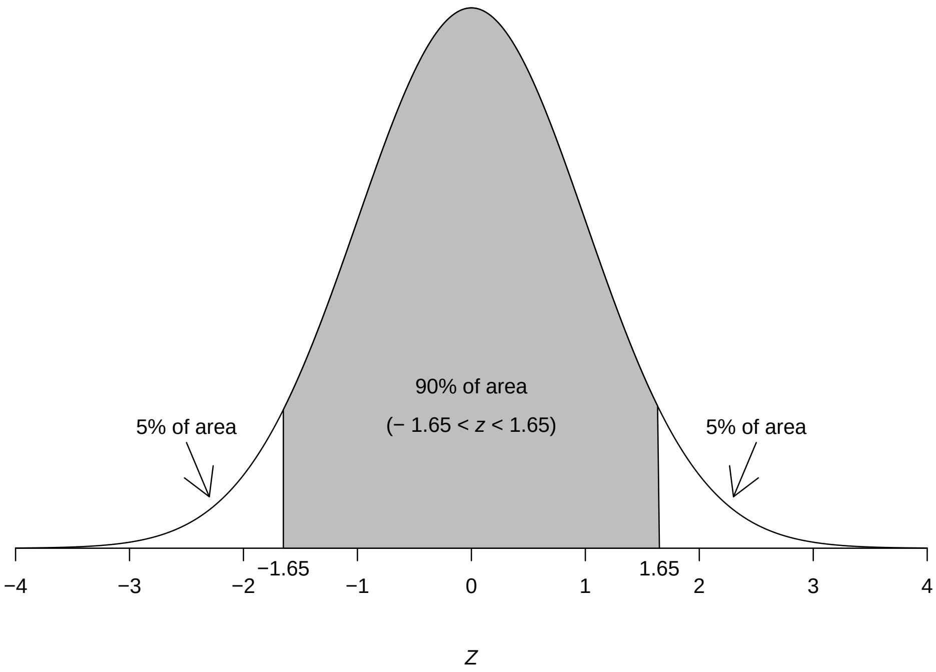
$$Pr\{Z < z\} = 0.90 \implies z = 1.28$$
$$Pr\{Z < z\} = 0.95 \implies z = 1.65$$
$$Pr\{Z < z\} = 0.90 \implies z = 2.33$$

But now we're interested in an interval, so instead of the *bottom* 90% (for example), we want the *middle* 90%. If we want the middle 90%, this means that we have 90% of our area in the middle (instead of the lower tail), and have 5% of our area in *each* tail. We are saying:

$$Pr\{z_1 < Z < z_2\} = 0.90 \implies z_1 = -1.65 \text{ and } z_2 = 1.65$$

Notice that if we want the middle 90% that implies $z_2 = 1.65$ which is the same value for z if we want the bottom 95%. Here's what we're doing in pictures:


95% of area
(z < 1.65)
5% of area
1.65
−4
−3
−2
−1
0
1
2
3
4
z


90% of area
(− 1.65 < z < 1.65)
5% of area
5% of area
−1.65
1.65
−4
−3
−2
−1
0
1
2
3
4
z

Similarly, we get:

$Pr\{z_1 < Z < z_2\} = 0.95 \implies z_1 = -1.96$ and $z_2 = 1.96$ (2 1/2% in each tail).
$Pr\{z_1 < Z < z_2\} = 0.99 \implies z_1 = -2.58$ and $z_2 = 2.58$ (1/2% in each tail).

So where does this leave us? We need to figure out how to take the above information and use it to construct our confidence interval. Suppose we took a sample and calculated $\bar{y}$. What we want to do is construct an interval around $\bar{y}$ in such a way so we can be (for example) 90% certain that this interval includes μ. Or to put it in a probability statement, we want:

$Pr\{y_1 < \bar{Y} < y_2\} = 0.90$

where y_1 and y_2 are chosen in such a way so that we're 90% certain they include μ.

As it turns out, we already know how to do this, although it's not quite obvious. We need to put all the pieces together. First we remember from above that (for a 90$ confidence interval):

$$Pr\{-1.65 < Z < 1.65\} = 0.90$$

Now we remember that we can convert from $\bar{y}$ to z using:

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

So we substitute this expression for Z in the equation above and get:

$$Pr\{-1.65 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.65\} = 0.90$$

Now we do some algebra (it's not difficult) and isolate μ in the middle:

$$Pr\left\{\bar{Y} - \frac{1.65\,\sigma}{\sqrt{n}} < \mu < \bar{Y} + \frac{1.65\,\sigma}{\sqrt{n}}\right\} = 0.90$$

And from this we can write the following expression that gives us a 90% confidence interval for μ:

$$\bar{Y} \pm 1.65\frac{\sigma}{\sqrt{n}}$$

Notice that this expression doesn't use μ (since that's what we're trying to guess at). But it does use σ. The obvious problem is that we don't know σ. Well, we could just substitute s for sigma to get:

$$\bar{Y} \pm 1.65\frac{s}{\sqrt{n}}$$

As hinted in the last chapter, this creates a problem because if we substitute s for σ in the equations above, we no longer have something that has a normal distribution. Let's look at this a bit further. If $\bar{Y} \sim N$ (i.e., $\bar{Y}$ has a normal distribution), then:

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N$$

The reason for this is that everything on the right hand side of this equation is a constant (remember, even though we don't know the value of μ and σ, they are constants). The only random variable on the right hand side of the equation is $\bar{Y}$. So if $\bar{Y} \sim N$, this implies that $Z \sim N$.

But now we have:

$$Z = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \quad \text{?? (see below)}$$

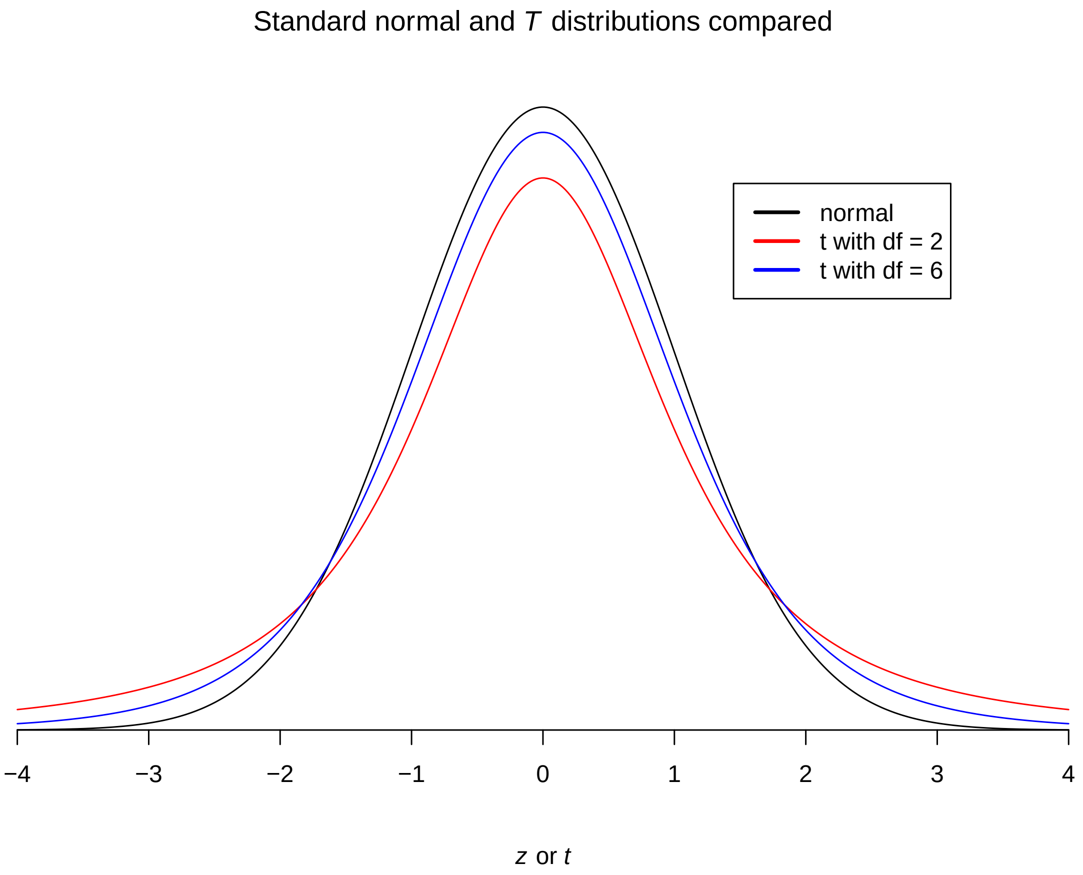
This time we have *two* random variables on the right hand side of the equation, $\bar{Y}$ and s. As before, we'll assume that $\bar{Y} \sim N$. But what about s? What kind of distribution does s have, and how does it affect Z?

It's not important for a class like this, but s has a chi (χ) distribution. The point is, it's not a normal distribution. So what do we get if we divide a random variable (e.g., $\bar{Y}$) that has a normal distribution by s? We do *not* get a normal distribution. Instead, we get something called a T distribution. We write:

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

The T distribution was developed by William Sealey Gosset. Gosset was very interested in the quality control of small samples (for brewing beer) at the Guinness brewing company in Ireland, and derived the T distribution while working for Guinness. He was not allowed to publish under his own name, so he choose the pseudonym "Student", and the T distribution has been known as Student's T ever since.

How can we use this T distribution? Before we figure that out, let's look at the T distribution and compare it to the normal distribution:



First, we see that the T distribution has a parameter we need to know about. This is given by the Greek letter ν (nu), and in this case is equal to something called the *degrees of freedom*, often abbreviated *df* (or sometimes *d.f.*).

As ν gets smaller, the tails of the T distribution get fatter. As ν gets larger, the tails of the T distribution get skinnier. As $\nu \to \infty$, the T distributions approaches the normal distribution.

To use the T distribution, we obviously need to know something about ν. For now, it turns out that this is fairly simple:

$$\nu = df = n - 1$$

In other words, ν is just our sample size minus 1. So let's get back to constructing our confidence intervals. We can't use the normal table, but we can use the $t-$table. If you look at the $t-$table, you'll see that it is arranged backwards from the normal table. That's because when we use the $t-$table we're almost always interested in reverse lookup. So the probabilities (or percentages) are arranged across the top of the table, and the $t-$values are in the body of the table.

Suppose you wanted to calculate the middle 90% of a T distribution with $\nu = df = 5$. Here's what you want in symbols:

$$Pr\{t_{5\%,5} < T < t_{95\%,5}\} = 0.90$$

(With the normal distribution we had $Pr\{-1.65 < Z < 1.65\} = 0.90$. We can rewrite this as $Pr\{z_{5\%} < Z < z_{95\%}\} = 0.90$, since $-1.65 = z_{5\%}$, and $1.65 = z_{95\%}$.)

Instead of -1.65 and 1.65 from the normal distribution, we now need to look in the $t-$table. We look across the top until we see 90%, then go down along the side until we see $df = 5$. We see 2.015. So now we can fill in our equation from above:

$$Pr\{-2.015 < T < 2.015\} = 0.90$$

Let's see if we can now use this to (finally!) construct a confidence interval. We need to modify the result from above. To construct a 90% CI, we can't use:

$$\bar{Y} \pm 1.65\frac{\sigma}{\sqrt{n}} = \bar{Y} \pm z_{90\%}\frac{\sigma}{\sqrt{n}}$$

Instead, we use (notice the $SE_{\bar{Y}}$):

$$\bar{Y} \pm t_{90\%,\nu}\,\frac{s}{\sqrt{n}} = \bar{Y} \pm t_{90\%,\nu}\,SE_{\bar{Y}}$$

Example: We take 6 ostrich eggs and weigh them and get an average weight of $\bar{y} = 1.4$ kg and standard deviation of $s = 0.175$ kg. We want to construct a 90% CI. We already know (very conveniently) that t for 90% and $df = 5$ is 2.015, so we proceed:

$$\bar{y} \pm t_{90\%,5}\frac{s}{\sqrt{n}} = 1.4 \pm t_{90\%,5}\frac{0.175}{\sqrt{6}} = 1.4 \pm (2.015)(0.07144) = 1.4 \pm 0.1440$$

Now we just have to finish this by subtraction and addition and we get:

$$90\%\,CI = (1.26, 1.54)$$

This is a *mathematical* interval, so make sure you write it correctly. For example, (1.54, 1.26) is ***not*** a mathematical interval (always put the smaller number first). Also, make sure you finish your calculations. 1.4 ± 0.1440 is also not an interval.

Before we continue, we need to clear up some possible confusion about subscripts.

First, we've been using $t_{90\%,\nu}$ and $t_{95\%,\nu}$ in the same equations. But we have actually been careful. When we say $Pr\{t_{5\%,5} < T < t_{95\%,5}\} = 0.90$, that means we want the middle 90%, which corresponds to 5% in the lower tail and 5% in the upper tail (5% in the upper tail corresponds to the 95$^{\text{th}}$ percentile). The t table is set up so that when we look up $t_{90\%}$ it gives us the values for t that put 5% in each tail (remembering to use the negative of the value in the table for the lower tail).

Second, a lot of textbooks use subscripts differently. Instead of doing $t_{90\%,\nu}$, they will use $t_{0.05,\nu}$. This is because putting 0.05 of the area in the upper tail corresponds with putting 0.10 (or 10%) of the area in both tails. If you look at our tables, for example, you'll see that the one sided probability of 0.05 corresponds with a CI of 90%. This will become a little more obvious when we do one sided tests, but for now, you may safely ignore this unless you're using a different text or different $t-$tables.

So back to CI's. Suppose for our ostrich egg example we now want to get a 95% CI. We already did most of the math, but now we need to look up the table value for $t_{95\%,5}$ and we get $t_{95\%,5} = 2.571$. So we have:

$$1.4 \pm (2.571)(0.07144) = 1.4 \pm 0.1837 \implies 95\%\, CI = (1.22, 1.58)$$

Or we could do a 99% CI:

$$1.4 \pm (4.032)(0.07144) \implies 99\%\, CI = (1.11, 1.69)$$

We can become more and more certain if we want, but as you can see, the CI keeps getting bigger and bigger. If we want to take this to the extreme, suppose we want to be 100% certain. The endpoints of the T distribution are at $-\infty$ and ∞. This is useless since this implies that a 100% CI is given by $(-\infty, \infty)$. This just tells us that μ exists (which we know), but is otherwise useless. In fact, usually we say that a 100% interval does not exist.

So what does a CI actually tell us? It tells us *how confident we are that the limits of our CI contains the true value of μ*. For our ostrich egg example, and a 90% CI we say: We are 90% confident that the true value of μ is between 1.26 and 1.54.

A CI is *not* a probability statement about μ. The following is *wrong*:

$$Pr\{20.9 < \mu < 42.5\} = 0.95 \qquad WRONG$$

Why? Because μ is a constant. It is not a random variable. We don't know what μ is, but it doesn't vary. Suppose, once again, that we do have secret knowledge and somehow did know that $\mu = 43.1$. What happens to the above probability? It *must* be 0:

$$Pr\{20.9 < \mu = 43.1 < 42.5\} = 0.00$$

In other words, a probability statement using μ must be 0 or 1. It is true or not true. This might seem an annoying technicality, but it is actually important to remember what μ represents.

Here's another way of looking at CI's. Suppose we construct 100 90% CI's (we take 100 samples, and for each sample we construct a CI). On average, 90 of our CI's will contain the true value of μ. Ten of them will *not* contain the true value of μ. The problem is, we don't know which of our CI's contain μ and which ones don't. If you construct just one CI you can only be 90% certain your CI contains μ. There's a 10% chance you'll miss μ. (Normally one would only ever take one sample to construct a CI - no one would take 100 samples and construct 100 CI's).

If you don't want to make a mistake 10% of the time, you can use a bigger CI. With a 95% CI you'll only make a mistake 5% of the time. With a 90% CI you'll only make a mistake 1% of the time. Problem is, the bigger you make your CI, the less information about μ your CI has (remember that a 100% CI is silly - and so is a 99.9999999% CI). When we learn about hypothesis testing we'll figure out another reason you don't want to make your CI too big.

Finally, we want to do one more example in constructing CI's. This one points out the limitations of using tables.

Example. We go out and collect 39 eggs from the American toad (*Anaxyrus americanus*). We want to construct a 99.9% CI for the diameter of our eggs (99.9% is a bit unusual, as we explained above). We find $\bar{y} = 1.5$ mm and $s = 0.25$ mm (our $\nu = df = 38$):

$$\bar{y} \pm t_{99.9\%,38}\frac{s}{\sqrt{n}} = 1.5 \pm t_{99.9\%,38}\frac{0.25}{\sqrt{39}} = 1.5 \pm t_{99.9\%,38}(0.04003)$$

Before we can finish this, we need to look up $t_{99.9\%,38}$ in our tables. Unfortunately there is no row for $df = 38$. So what do we do?

The quick answer (we'll explain below) is that we *always* use the row with *less* degrees of freedom than we have. We always round down. In this case we use the row for 35 df.

So we go into our t table and find $t_{99.9\%,35} = 3.591$ And we get:

$$1.5 \pm (3.591)(0.04003) \implies CI = (1.3562, 1.6438)$$

And we are 99.9% confident that the true average diameter (μ) of toad eggs is between 1.4089 and 1.5911mm.

So why do we always round down? Let's take a look at the confidence intervals for both 35 df and 40 df:

For 35 *df* we have the CI we calculated above: (1.3562,1.6438).

For 40 *df* we get $t_{99.9\%,40} = 3.551$ which gives us a CI of (1.3579,1.6421).

While it isn't a big difference, notice that the interval for $df = 40$ is smaller than the one for $df = 35$. In other words, the interval for 40 *df* is *too small*, while the interval for 35 *df* is *too big*. We always want to be *at least* 99.9% sure. Since the interval for 40 *df* is to small, we can't say this.

We know the interval for 35 *df* is too big, but that's okay since we're at least 99.9% sure our interval includes μ.

Finally, we could use interpolation, but interpolation is not used much these days as it's a bit of a pain and most software will give us the exact values we need in any case. We can easily ask R and get $t_{99.9\%,38} = 3.566$ (notice this value is between the two values for t given above).