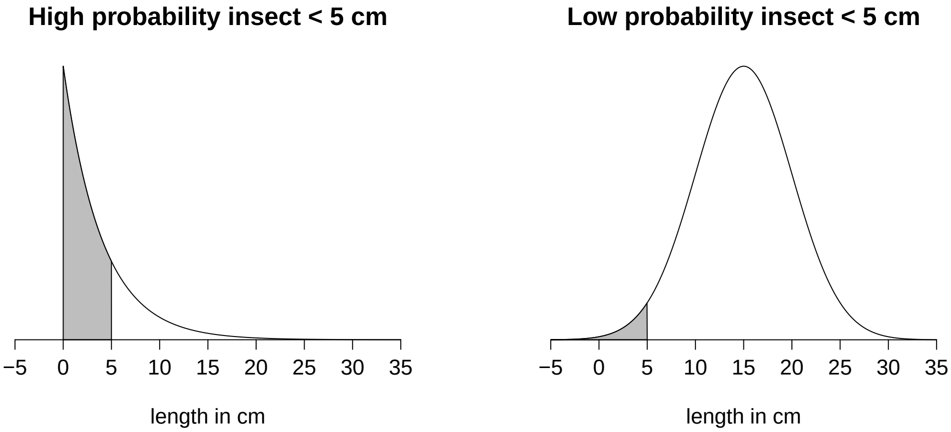
Distributions

Distributions and Probability

When we look at a random variable such as Y, one of the first things we want to know is *what is it's distribution?* In other words, if we have a large number of Y's (say, 50 measurements of tree diameters in cm), we want to know what the shape of the distribution is. As we learned earlier, one easy way to do this is to make a histogram of our data.

But why do we want to know the shape of our random variable? Because how we proceed with our analysis is depends on the shape of our distribution. This is because the shape of our distribution directly affects the probability of various outcomes or results. As an example, let's look at the length of a hypothetical insect and plot its distribution. The first time we'll assume a highly skewed distribution and the second time we'll assume a typical bell shaped distribution.



Obviously, the probability of our insect being less than 5 cm depends a lot on the shape of our distribution.

Often, the shape of the distribution can be determined by what we're interested in. Suppose, for example, we toss a coin 10 times and $Y = \textit{number of heads}$. If we know this, then we also know that Y will have a binomial distribution (see the previous chapter on probability). We can write:

$$Y \sim Binomial$$

The ”$\sim$” symbol means *distributed as*, so this means our random variable, Y, is distributed as a binomial distribution. Often we also put in the *parameters* of our distribution into this same expression. We'll explain more about parameters below, but notice that the sample size (n) and probability of success (p) are the parameters of a binomial distribution. In this case we have $n = 10$ and $p = 0.5$, so we can re-write our expression as follows:

$$Y \sim Binomial(10, 0.5)$$

or, simpler:

$$Y \sim B(10, 0.5)$$

Finally, we should mention that this is the theoretical distribution of our random variable (Y). If we actually toss a coin 10 times and to this many, many times (say 1,000 times), then our *actual* distribution may eventually look like our theoretical distribution (if our coin is fair).

More on the binomial distribution

Let's briefly review the binomial distribution that we first introduced when we looked at probability. Here's the equation again:

$$\binom{n}{y} p^y (1-p)^{n-y}$$

So what makes this a distribution? Several reasons (this breakdown may vary a bit in different texts):

1. We can use this equation to calculate the probability of any (or all) possible outcome(s).
2. All possible outcomes add up to 1 (the probability of *something* happening is 1).
3. The actual shape of the distribution is determined by the *parameters* of the distribution.

Let's explain these in a bit more detail using our coin example (10 tosses with a fair coin). We can use our binomial distribution formula to calculate all possible outcomes. For example, the probability of 0 tails in 10 tosses would be:

$$\binom{10}{0} 0.5^0 (1-0.5)^{10-0} = 0.00098$$

The probability of 1 tail in 10 tosses would be:

$$\binom{10}{1} 0.5^1 (1-0.5)^{10-1} = 0.00977$$

The probability of 2 tails in 10 tosses (which, incidentally, is equal to the probability of 8 tails in 10 tosses (only true if $p = 0.5$), which we calculated in the chapter on probability) would be:

$$\binom{10}{2}0.5^2(1-0.5)^{10-2} = 0.04395$$

We could go on and do all 11 possible outcomes, but let's just list the answers in a table:

Heads	Tails	Probability
10	0	0.00098
9	1	0.00977
8	2	0.04395
7	3	0.11719
6	4	0.20508
5	5	0.24609
4	6	0.20508
3	7	0.11719
2	8	0.04395
1	9	0.00977
0	10	0.00098
Sum:		**1.00000**

A summary like this that lists all possible outcomes can be very useful. For example, we can now easily calculate the probability that $Y = 0$, 1, or 2 heads:

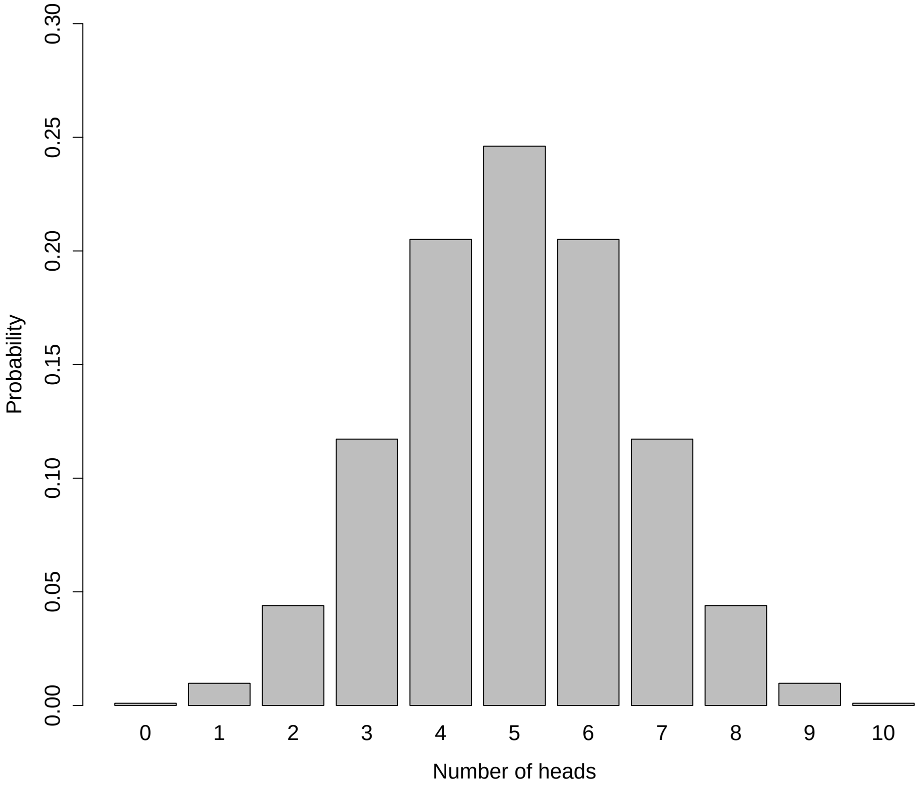
$$Pr\{0 \leq Y \leq 2\} = 0.00098 + 0.00977 + 0.04395 = 0.05470$$

Also notice that all possible outcomes add to 1.0:

$$Pr\{0 \leq Y \leq 10\} = 1.0$$

This is a *very* important point that should be obvious: if we toss a coin, *something* has to happen. Since the above table lists every possible outcome, these outcomes add to 1.

So let's take a look at our (theoretical) distribution and see what it looks like. To do this, we simply plot each probability as the height on a bar graph:



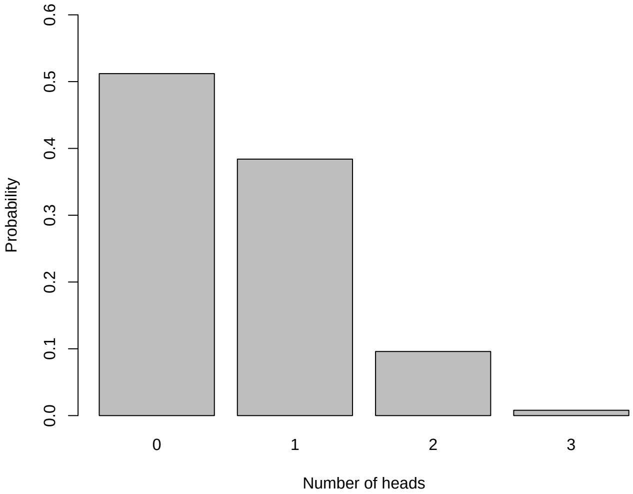
This gives us a visual representation of the distribution of our random variable.

So what about the parameters? *Parameters determine what our distribution looks like.* In our coin tossing example our parameters were $n = 10$, and $p = 0.5$. So what happens if we change our parameters? The *shape* of our distribution changes.

Suppose we had an unfair coin. Somehow we've managed to rig our coin so it comes up heads only 20% of the time. That implies our $p = 0.2$. Now let's assume we toss it three times (n =3). What does this do to the shape of our distribution? Let's calculate all possible outcomes first, just like we did above (remember Y = number of heads):

Y	Probability
0	0.512
1	0.384
2	0.096
3	0.008
Sum:	**1.00000**

If you don't know how we got the probabilities in the table above, you should review the earlier parts of this chapter. But let's plot this just like we did before:



And we notice that the shape (and number of bars) has changed considerably. In other words, *the parameters determine what our distribution looks like!*

Summarizing distributions so far

Based on what we learned about the binomial distribution, we can now remind ourselves of two properties of distributions:

1. The shape of a distribution varies based on the parameters.
2. All possible outcomes must add up to 1.

Let's investigate property (2) a little more. If we have a discrete distribution (if Y is discrete), then it's easy to see how to add up all possible outcomes. Again, using the binomial as an example, what we are saying is:

$$\sum_{y=0}^{n} \binom{n}{y} p^y (1-p)^{n-y} = 1.0$$

This principle is not true just for the binomial, but for *any* distribution. If our distribution is discrete, we can just rewrite our equation as follows:

$$\sum_{y=0}^{n} f(y) = 1.0$$

Where $f(y)$ is our discrete distribution (e.g., the binomial).

But what do we do if our distribution is continuous? How can we possible add up all possible values? Let's think about this for a moment by illustrating a simple probability for a continuous variable. We'll let $Y = \textit{height of a woman.}$ So let's figure out the following:

$$Pr\{Y = 65 \text{ inches}\} = ??$$

This might seem like a simple calculation (and it is), but the answer isn't obvious. Let's rewrite this to make the answer a bit more obvious:

$$Pr\{Y = 65 \text{ inches}\} = Pr\{Y = 65.0000000000000000000000000... \text{ inches}\} = 0.0$$

In other words, no one is *exactly* 65 inches tall. So how do we add up all possible outcomes, if the probability of any particular event is 0? Unfortunately, the answer requires calculus.

You are not responsible for calculus, but let's investigate this just a bit further. Using calculus, mathematicians (and statisticians!) can add up a bunch of infinitely small things like the probability of someone being between 65 and 66 inches (the probability of any one outcome is 0, but amazingly we can still add these up!). What we are saying is something like the following:

$$\int_{-\infty}^{\infty} f(y)dy = 1.0$$

Where this time $f(y)$ represents our continuous distribution. Note that if you replace the $\int$ symbol with the $\sum$ symbol and ignore the dy, you essentially have the same thing we did above. In fact, historically the $\int$ symbol *means* "sum".

So to summarize, and temporarily get away from calculus, we can use calculus to add up a sequence of infinitely small things. If we use this to add up all possible outcomes, we *must* get 1. Or, as a simple example, the probability that a particular woman has a height (any height) is 1.0 (she *must* have a height, or she wouldn't exist). So any distribution, discrete or continuous will give us 1 if we add up all possible outcomes.

Introducing the normal distribution

The normal distribution is without a doubt the most important distribution in statistics. Some of the reasons for that won't be apparent for a while, but have to do with the properties of random variables.

As an aside, the normal distribution is also knows as the Gaussian distribution after Carl Friedrich Gauß (1777 - 1855), who did a lot of work with it. Gauß was one of the most famous mathematicians of all time and made major contributions to number theory, statistics, geometry, linear algebra and many other fields. In fact, the Germans used his picture on the 10DM bill and even put the equation (yes, the equation!) for the normal (or Gaussian) distribution on the bill right next to his picture (the DM was the old German currency before the introduction of the Euro). If you're interested in Gauß, you should check out his page on Wikipedia.

So what is the normal distribution? It is given by the following:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}$$

So now you know everything, right? Seriously, let's give a few explanations and then do some examples. The parameters of the normal distribution are μ and σ, and you can see both of them in the expression above. There are two other constants you should be aware of. The first is π, which, as you would expect is the ratio of a circle's circumference to its diameter and is given by $\pi = 3.14159265$...etc. The other constant is e, which is the base of the natural log and is given as $e = 2.71828182$...etc. Like π, e is an irrational number.

So why is this distribution so important? Two main reasons (for us):

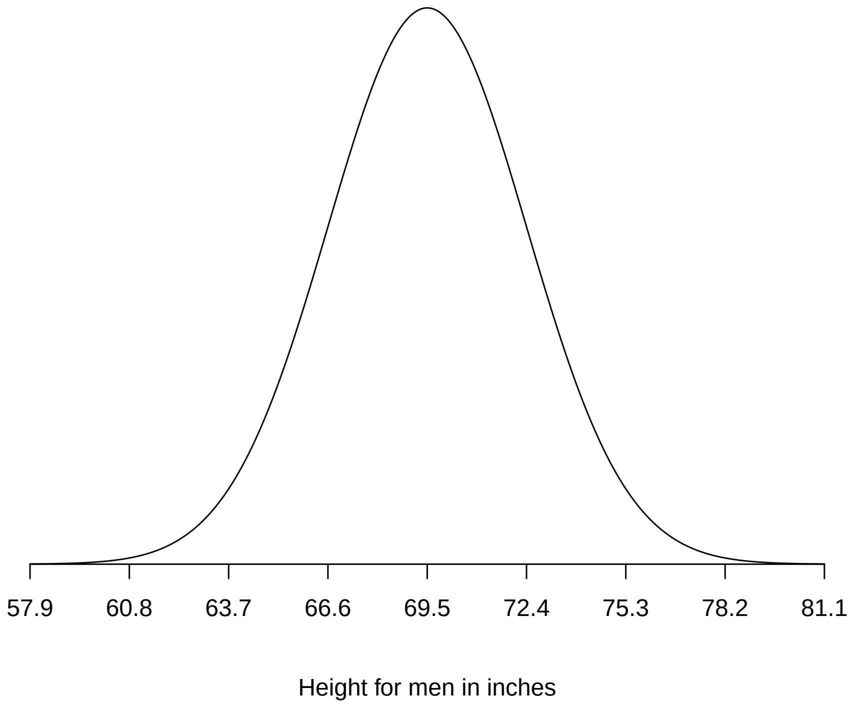
1. Because many things in biology have a distribution that is approximately normal.

2. Because of the Central Limit Theorem (CLT). This theorem let's us calculate probabilities using the normal distribution even if Y is not distributed normally. We will learn more about the CLT later, but it is safe to say it is one of the most important theorems in statistics.

If you have a graphical calculator (or software like R), you can easily plot this equation and you will get pictures similar to the ones below. Let's take a look at the normal distribution. We'll plot the distribution of heights for adult men. Before we can do this, we need to know μ and σ. Obviously, as explained earlier when we discussed samples and populations, we *can't* know μ and σ, but let's pretend that we do, and somehow have secret knowledge:

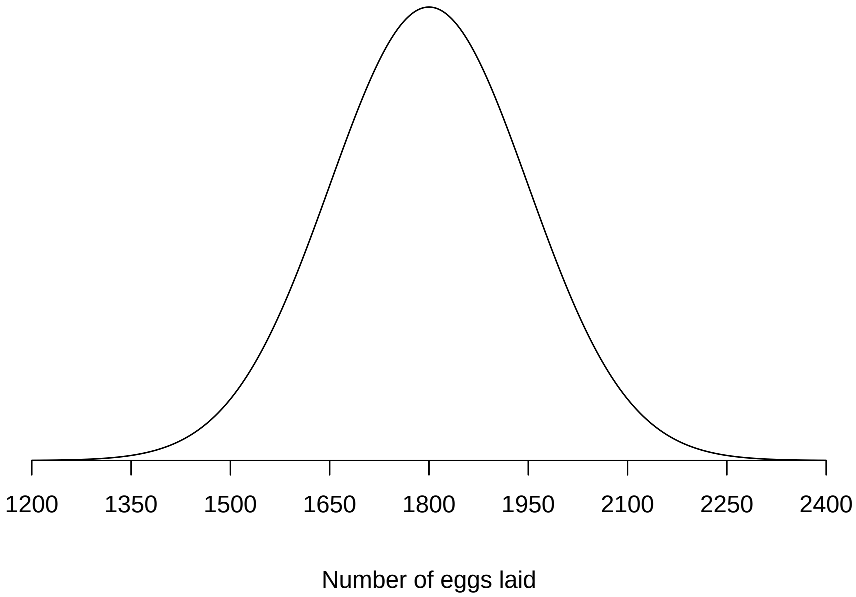
$$\mu = 69.5 \text{ inches and } \sigma = 2.9 \text{ inches}$$

Now we can use these parameters in our equation for the normal distribution and we get the following picture (note the scale on the x - axis):



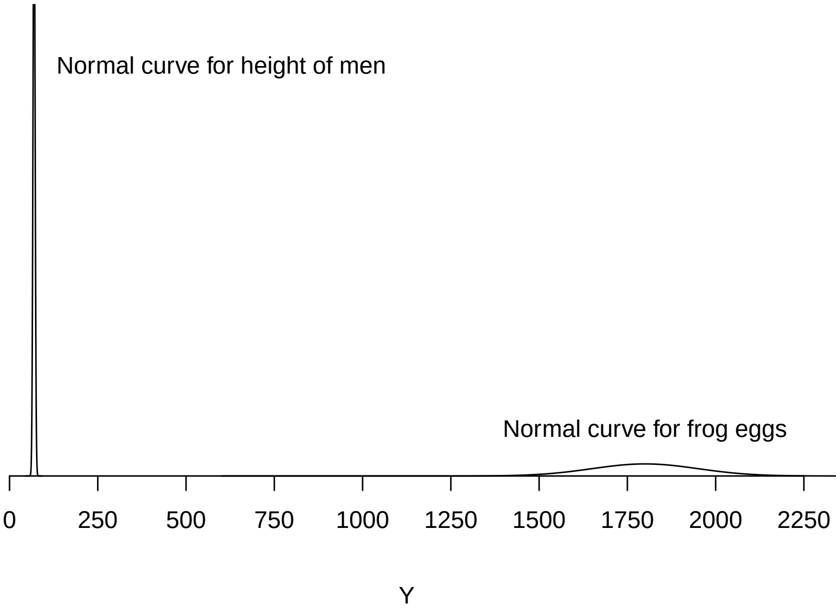
We notice that the curve reaches its peak (maximum value) at $\mu = 69.5$. Also notice that the curve changes direction (has an inflection point) at $\mu \pm \sigma$, or in this case at 69.5 ± 2.9. Finally, notice that the ends of the curve go from $-\infty$ to $+\infty$.

Let's try this again, but this time we'll look at the number of eggs laid by Gray Tree Frogs (*Hyla versicolor*). Again, let's pretend we have secret knowledge and know that $\mu = 1,800$ and $\sigma = 150$:



This plot looks identical to the one for the height of men (are they the same?). Let's make some comments on the normal distribution.

First, if you look at the scale of the two normal curves, they are obviously not the same. The first is centered at 69.5, the second at 1,800. Also, the first is considerably less wide than the second. Let's visualize this by plotting them both on the same graph:



As you can see, the two normal curves actually look rather different if drawn to the same scale. As should be obvious, the *parameters* of the normal curve are μ and σ: the two numbers we changed for the two normal curves above.

Conveniently, the mean and standard deviation of a normal curve are actually the parameters (μ and σ). The mean (μ) determines the exact location of the normal distribution on the $x-$ axis. The standard deviation (σ) determines how spread out the normal distribution is (this should be obvious looking at the graph above).

So how does this help us in calculating probabilities? As implied above, the area under the normal curve has to add to 1 (this is a fundamental property of a distibution). What we are saying (using calculus) is:

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2} dy = 1.0$$

While we can't use a continuous distributionlie this to calculate the probability that Y is equal to any one value, we *can* use it to calculate the probability that Y is greater or less than any particular value. For example, we can calculate the probability that a particular gray tree frog lays less than 1700 eggs:

$$Pr\{Y < 1700 \text{ eggs}\}$$

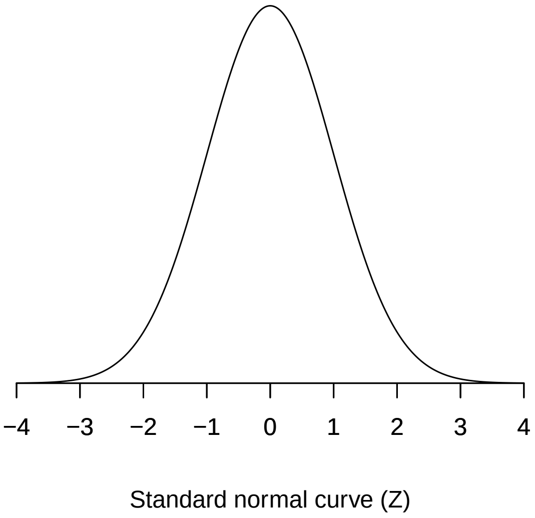
If you do know calculus, you might think the way to calculate this probablity is as follows:

$$\int_{-\infty}^{1700} \frac{1}{150\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-1800}{150})^2} dy$$

Where we plugged in $\sigma = 150$ and $\mu = 1800$. Unfortunately, this equation isn't solveable analytically, so we need to resort to other techniques. What we do is use a table to look up this probability (or letting R do this for us).

But before we can do this, we have one more problem to solve. We illustrated two different normal curves above. How many different normal curves are there? If you think about this for a minute, we can't possibly make a table for every possible normal curve.

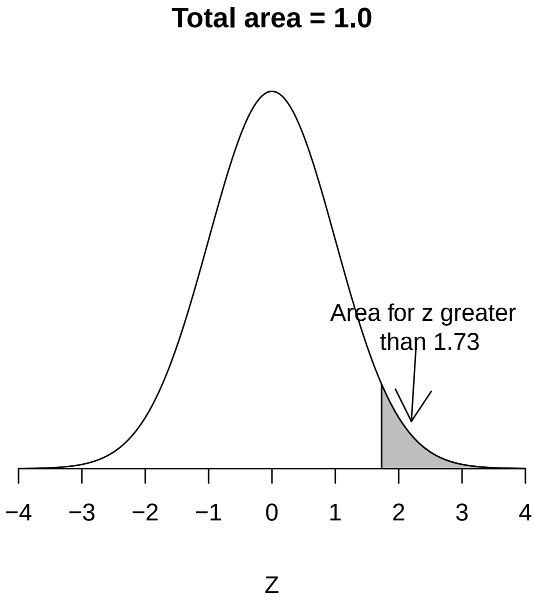
Instead, we will use just one normal curve to look up our probilities and convert the values of our curve to this normal curve (see below). The curve we choose is called the *standard normal curve* and has $\mu = 0$ and $\sigma = 1$.



Note that we also use Z instead of Y for this distribution. We say $Z \sim N(0, 1)$.

So how do we use this to calculate probabilities? The values for this curve are tabulated in the standard normal table. Here's how to use it:

Suppose we want to find the probability that Z is greater than 1.73 (we want $Pr\{Z > 1.73\}$). First let's draw a picture of what we want:



The probability (which is the area above) should be fairly small. Our normal distribution table gives the you the area *less* than a particular value of Z. The side gives the first two digits of z, and the top gives the third digit of z. So to get the area less than Z = 1.73 (remember we want greater than) we go into the table and look for "1.7" on the side of the table and 0.03 across the top of the table. Then we pick out our area (= probability) and get 0.9582. In other words:

$$Pr\{Z < 1.73\} = 0.9482$$

We're almost done. If you remember, we want $Pr\{Z < 1.73\}$, not $Pr\{Z < 1.73\}$. Since the area under the curve is 1, we can simply subtract to get our answer:

$$Pr\{Z > 1.73\} = 1 - Pr\{Z < 1.73\} = 1 - 0.9482 = 0.0418$$

And we have our answer.

There is a slightly easier way to do this, particulary once you get comfortable with this idea. We notice that the curve is completely symmetrical around 0. We can take advantage of this:

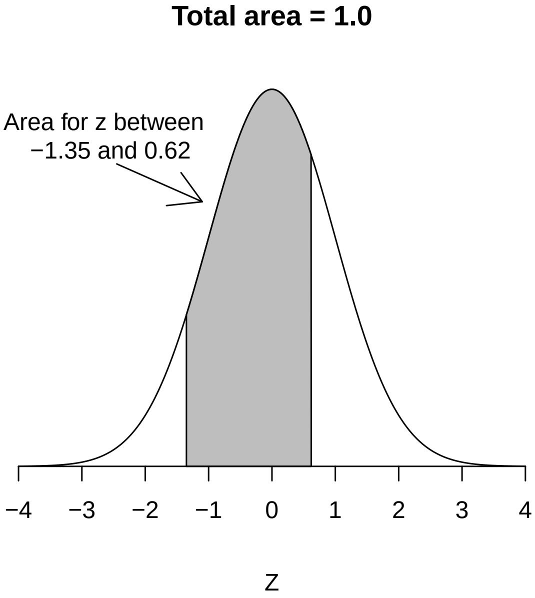
1. Change the sign of the z we're interested in. In our example we would use -1.73 instead of 1.73.

2. Change the inequality and then simply look up $Pr\{Z < -1.73\}$ in the normal table.

And we find that $Pr\{Z < -1.73\} = 0.0418 \quad (= Pr\{Z > 1.73\})$

This always works due to the symmetry of the normal curve, which is why some text will only give you one half of the normal table. If you find this confusing, just stick with the method outlined above.

Let's try another example and try to find the probability that Z is between -1.35 and 0.62 or $Pr\{-1.35 < Z < 0.62\}$. Here's what we want this time:



To get the area in between, we will need to subtract two ares. First, let's get the area less than 0.62:

$$Pr\{Z < 0.62\} = 0.7324$$

This gives us *all* the area for z less than 0.62, but that's not what we want.

So let's now get the area less than -1.35:

$$Pr\{Z < -1.35\} = 0.0885$$

And this gives us everything below -1.35.

To get what we want, we subtract this area from the previous area, and that will give us the area we want:

$$Pr\{-1.35 < Z < 0.62\} = 0.7324 - 0.0885 = 0.6439$$

If this is a little confusing, you might want to go through this example again, and look at the last example using elephant weights below.

The standard normal curve lets us calculate probabilities by looking up values in the normal table. But how do we use this to calculate probabilities for Y? How do we get the probability of a gray tree frog laying less than 1700 eggs? We need to convert Y into Z, and then we can use the standard normal table.

Here's how to do it:

$$z = \frac{y - \mu}{\sigma}$$

This transforms our Y (with whatever μ and σ) into Z (which has $\mu = 0$ and $\sigma = 1$). Let's try it out with our gray tree frog example:

$$z = \frac{1700 - 1800}{150} = -0.667$$

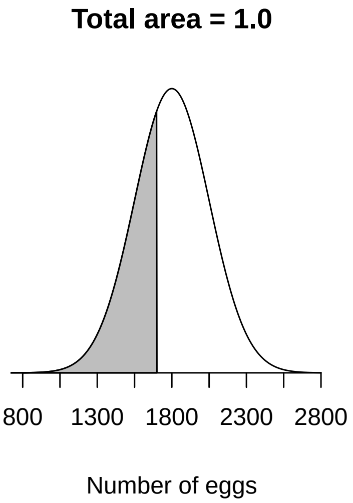
This means that:

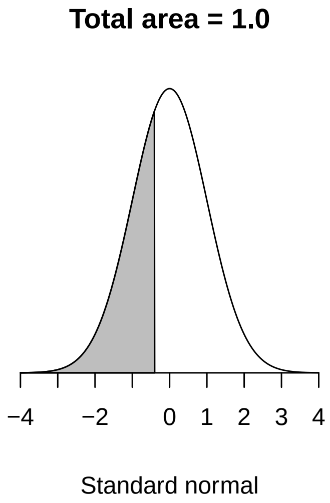
$$Pr\{Y < 1700\} \equiv Pr\{Z < -0.667\}$$

The symbol $\equiv$ means *exactly equivalent to.* So now we just have to look up $Pr\{Z < -0.667\}$ in our normal table, and we get:

$$Pr\{Z < -0.667\} = 0.2514$$

And so the probability that a gray tree frog will lay less than 1700 eggs is 0.2514. Here's a picture of what we did:



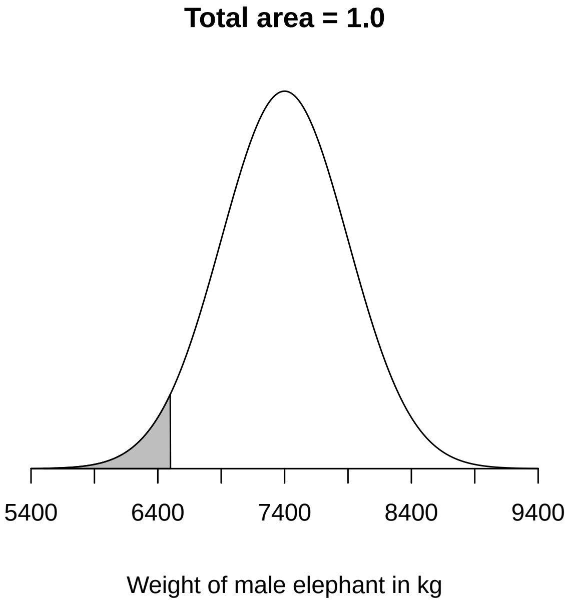


The area in the two graphs above is identical, which is why we can use the standard normal distribution to look up our probablities.

Let's try a few more examples using African Elephants (*Loxodonta africana*). Somehow we know that for male elephants, $\mu = 7,400$kg and $\sigma = 500$kg.

Example 1: We want to find the probability an elephant weight less than 6,500kg:

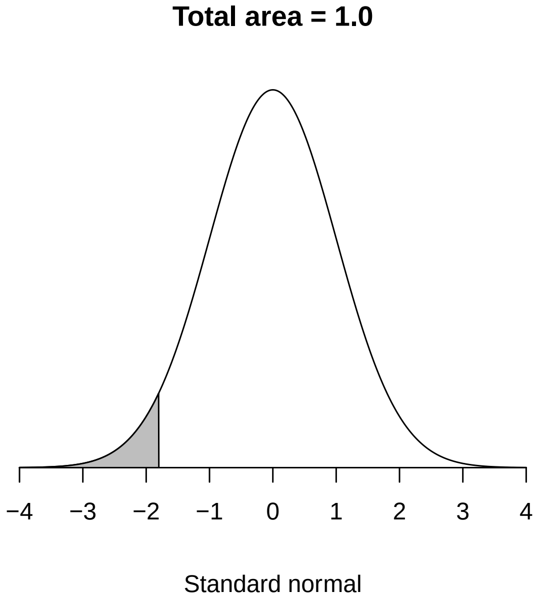
First, let's draw a picture of what we want using elephant weights:



Now let's convert our Y to Z:

$$z = \frac{y-\mu}{\sigma} = \frac{6,500 - 7,400}{500} = -1.8$$

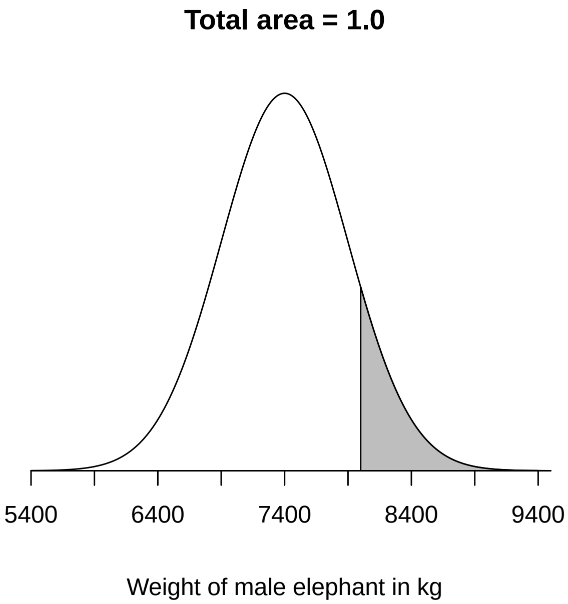
And we can look at what we want on the standard normal curve (notice the areas are identical, as they should be!):



Finally, we look up -1.8 inour table and get 0.0359. So we can conclude:

$$Pr\{Y < 6,500\} = Pr\{Z < -1.84\} = 0.0359$$

Example 2: We want to find the probability an individual elephant weight more than 8,000kg. Again, let's start with a picture (only one this time):



Now let's convert our Y to Z as in the last example:

$$z = \frac{y - \mu}{\sigma} = \frac{8,000 - 7,400}{500} = 1.2$$

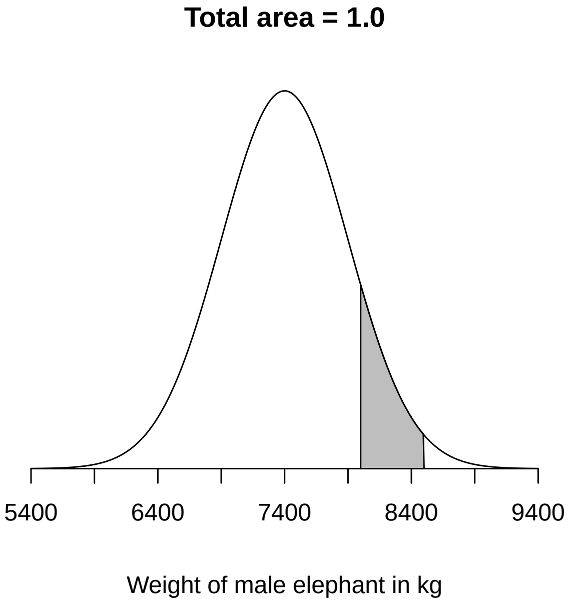
We look up $Pr\{Z > 1.2\} = 0.8849$, and finally we do:

$$Pr\{Y > 8,000\} = Pr\{Z > 1.2\} = 1 - 0.8849 = 0.1151$$

Or we could have noted that $Pr\{Z > 1.2\} = Pr\{Z < -1.2\}$ and looked up:

$$Pr\{Z < -1.2\} = 0.1151$$

Example 3: Finally, let's get the probability an individual elephant weight more between 8,000kg and 8,500 kg. Here's our picture this time:



This time we need two values of Z:

$$z_1 = \frac{8,000 - 7,400}{500} = 1.2$$

$$z_2 = \frac{8,500 - 7,400}{500} = 2.2$$

This implies that $Pr\{8,000 < 8,500\} = Pr\{1.2 < Z < 2.2\}$. To get our answer, we look up $Pr\{Z < 2.2\} = 0.9861$. This gives us all of the area less than 2.2. However, we don't want all of this area, we want the area between 1.2 and 2.2, so we need to *subtract* the area less than 1.2. In other words we look up $Pr\{Z < 1.2\} = 0.8849$. Now we can do our subtraction and conclude:

$$Pr\{8,000 < Y < 8,500\} = Pr\{1.2 < Z < 2.2\} = 0.9861 - 0.8849 = 0.1012$$

Now that we have figured out how to calculate (look up) probabilities using the normal distribution, we need to do one more thing. Suppose we have a particular probability in mind, and want to figure out the value of Y that goes with that probability? This may sound a bit confusing, so let's illustrate this using our elephants.

I want to know the weight that corresponds to 90% of my elephants, or to put it another way, 90% of elephants weigh less than __? We want to fill in that blank. If we put this into a probability statement, here is what we're after:

$$Pr\{Y < y\} = 0.90$$

We want to find the y that makes the above statement true. What value of y is equivalent to 90% of the area? We also call these values percentiles. For example:

> The 90^{th} percentile has 90% of the values of Y below y.
>
> The 64^{th} percentile has 64% of the values of Y below y.
>
> (This might be a good time to remember that Y is a random variable and y is an actual value).

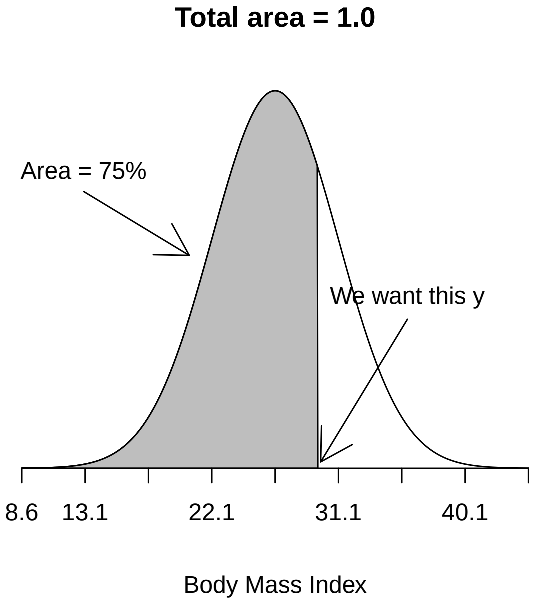
Let's motivate this a bit further by looking at an example from academic testing. Many colleges require students to take the SAT or ACT before they are admitted. These tests (supposedly) measure how well prepared you are for college level work. In addition to a score, these tests will also give you a percentile. These percentiles tell you how you compared to everyone else taking the same exam.

For example, you score 73% on a part of the test. This means you did *better* than 73% of the people taking this test (for that part). Your score, whatever it was, corresponds to the 73^{rd} percentile.

Many testing agencies use percentiles because it is informative and also makes variations in the test from one version to the next comparable. In other words, if you scored in the 73^{rd} percentile this year and in the 73^{rd} percentile last year, your performance was similar compared to everyone else, regardless of what your scores were (your scores could be different, but your percentiles are the same).

So what does all this mean for us? We need to learn how to calculate (look up) the percentiles. This is often called *reverse lookup*. Let's do an example using the body mass index (BMI). Yes, the body mass index has problems, particularly for individuals (e.g., it's not very useful for many athletes), but it's quick, easy, and does reasonably well when looking at large groups of people.

According to the CDC, the average BMI for men in the United States is about 26.6, and the standard deviation is about 4.5. Let's try to find the 75^{th} percentile for men assuming the CDC figures actually represent the population parameters (so $\mu = 26.6$ and $\sigma = 4.5$). First let's draw a picture of what we want:



So here is how to do it. We remember that our normal distribution table gives us the area (= probability) below the number (z) that we look up. Normally we would use our z to look up the probability. Now we do things backwards. In other words, Now we want to find z in $Pr\{Z < z\} = 0.75$ (up until now we calculate z and look up the probability (0.75, in this case)).

To do this, We go *into* the body of the table (not along the sides or top) and look for the number that's closest to 0.75. This turns out to be 0.7486 (a little closer to 0.75 than the next number, 0.7517). Now that we have found 0.7486, we look up z: we move from 0.7486 to the side and fine 0.6, we go to the top and find 0.07, so we now $z = 0.67$. We now know that:

$$Pr\{Z < 0.67\} \approx 0.75$$

We will not worry about extrapolation in our class - instead we will just get the closest number.

Now all we need to do is to convert our z back to y. We just rearrange our equation for z and get:

$$z = \frac{y - \mu}{\sigma} \quad \Longrightarrow \quad y = z\sigma + \mu$$

So we can do:

$$y = 0.67(4.5) + 26.6 = 29.615$$

And we can say that the 75$^{\text{th}}$ percentile for BMI in men is 29.615, or 75% of men have a BMI that is less than 29.615.

Finally, we should remember that statistics is about more than just numbers and think about our results just a bit. If we realize that a BMI of 25 or higher is considered overweight, and 30 or higher is obese, what do these numbers tell us about men in the U.S.?