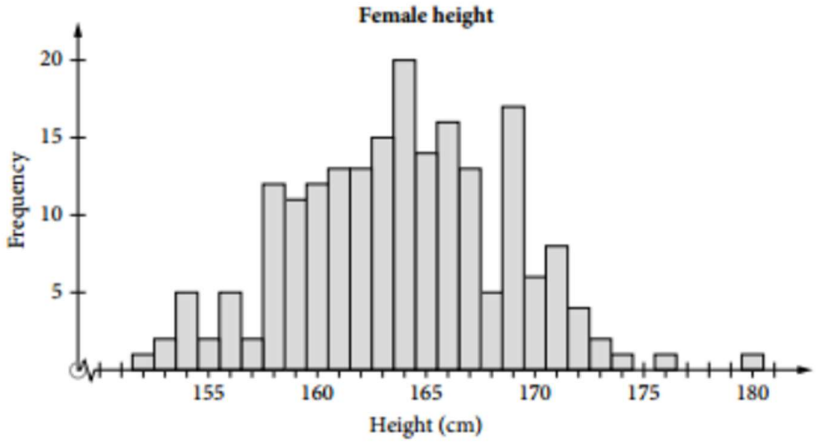
THE NORMAL DISTRIBUTION

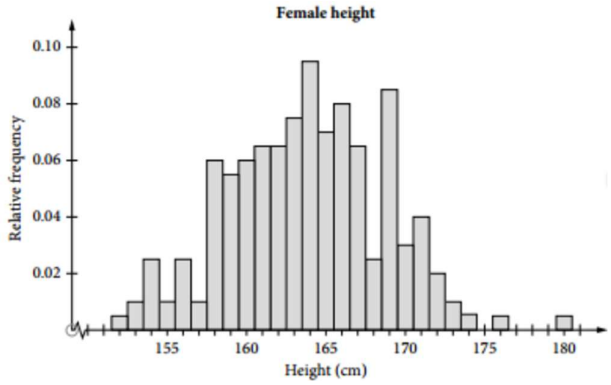
The following histogram shows the height of 200 randomly selected Australian women, measured in cm accurate to one decimal place.



Can you find a function that will model the data, just like Dirk did with the train data in the previous module?

The total area of the rectangles in the histogram is 200, the size of the sample, so the first thing that needs to be done is to make the sum of the areas of the rectangles equal to 1.

To do this, divide the vertical scale values by 200. As the bin width (the width of the columns in the histogram) is set to 1, this means that the area of each bin represents the probability of a woman having a height in that 1 cm range. More importantly, the total area of the bins is now 1, so the situation is like a probability density function.

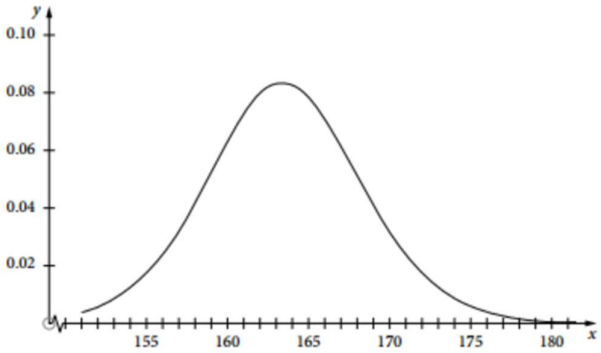


From the raw data, the mean and standard deviation can be calculated:

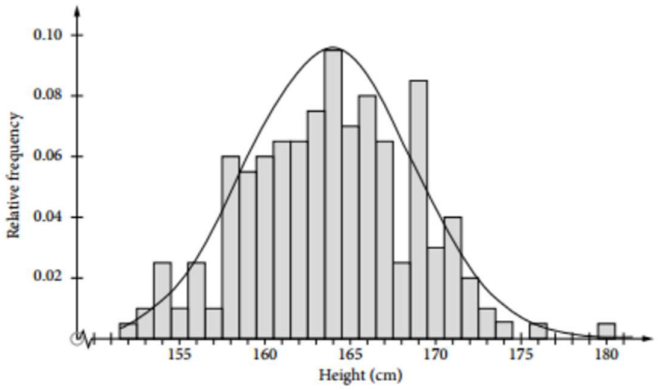
Mean = 163.8085

Standard deviation = 4.858 855

What sort of function might be suitable as a model for this data? The shape appears similar to a parabola, but the ends seem to drop away too quickly for that to be a successful model. Perhaps a shape such as the one shown below might be suitable.

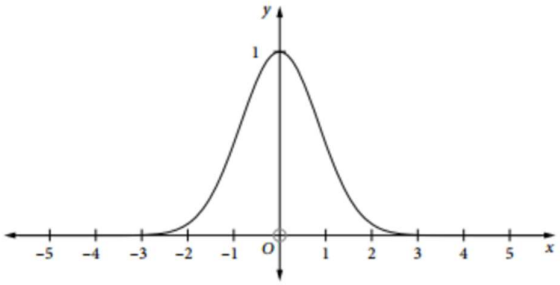


If the two images are put together it can be seen that this function certainly has some potential as a model for the distribution of heights.



This matches the general shape quite well, although as with all models, it is not perfect. What function might produce this graph? The way the graph falls off suggests an exponential function of some kind, but unlike a typical exponential it falls off on both sides. Instead, try $f(x) = 2^{-x^2}$.

This shape looks like a good match, but is it a pdf? The values of the function are non-negative, so it satisfies that condition for a pdf. The graph almost touches the x-axis at -3 and 3, so the area could be approximated using a triangle with a base of 6 and a height of 1. That gives an area of 3, which is too large a value. The graph of the function will need to be adjusted, possibly by using a vertical dilation, for example $f(x) = k \times 2^{-x^2}$ for some suitable value of k. This can be determined by using a definite integral to obtain a more accurate approximation.

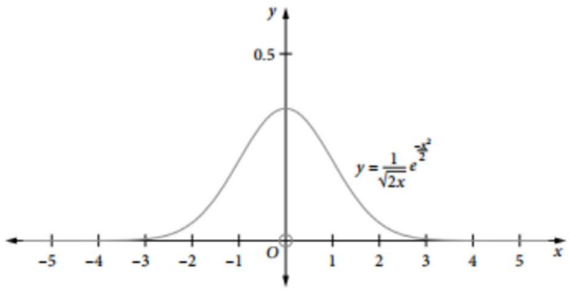


Note that while the domain of the function is $\mathbb{R}$ (real numbers), the value of the function is increasingly close to 0 to the left and right of the graph, for example, $x < -3$ and $x > 3$.

Using a suitable definite integral to find an approximation for the area beneath the graph, therefore find the value of $\int_{-3}^{3} 2^{-x^2}\,dx$. The integration of this function is not covered in this course, but if you have access to graphing software you can use that to evaluate this integral. The integral could also be approximated using the trapezoidal rule.

Hence $\int_{-3}^{3} 2^{-x^2}\,dx = 2.12805676757 \approx 2.12806$.

This shows that using $k = 2.12806$, and hence $f(x) = \frac{1}{2.12806} 2^{-x^2}$, will produce a pdf with the desired shape. However, it will also need to be translated and dilated to match the data presented at the beginning. Mathematicians have found that a base of e rather than 2 works well, and in this case instead of $\frac{1}{2.12806}$ the value $\frac{1}{\sqrt{2\pi}}$ should be used. This gives the function with domain $\mathbb{R}$ and rule $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$, defined using constants you are already familiar with. This graph is shown at right and has the same general shape.



Although it looks like the graph meets the x-axis, it is asymptotic to the x-axis in both directions. That is, as $x \to \infty, f(x) \to 0$ and as $x \to -\infty, f(x) \to 0$ where $f(x) > 0$ for all $x \in \mathbb{R}$.

You can evaluate $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\,dx$ using graphing software.

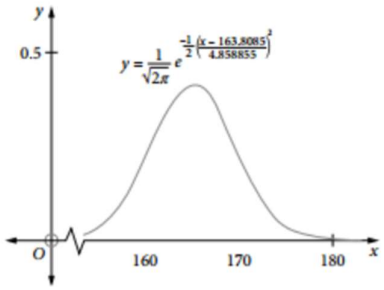
THE NORMAL DISTRIBUTION

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\,dx = \int_{-100}^{100} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\,dx = 1$$

(The limits on the integral have been changed as ∞ cannot be used by all software, but [−100,100] is a close enough approximation to infinity; −50 to 50 also gives a value of 1.)

You now have a pdf, but you still need to translate and dilate this to match your data. From your previous work of transformations of graphs of functions, you should recall that a horizontal translation by h units will transform the graph of $y = f(x)$ onto the graph of $y = f(x - h)$. You should be able to see that the required horizontal translation is equal to the mean; however, a horizontal dilation will also be required. What value might be related to the horizontal dilation? The effect of the horizontal dilation is to spread out the values. The standard deviation is a measure of the spread of the data, so that is a value worth trying.

In combination, a horizontal translation by h units combined with a horizontal dilation by b units will transform the graph of $y = f(x)$ onto the graph of $y = f\left(\frac{x-h}{b}\right)$. In this case, the value of h is given by the mean of your data, 163.8085, and the value of b is given by the standard deviation, 4.858855. The function obtained is $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-163.8085}{4.858855}\right)^2}$

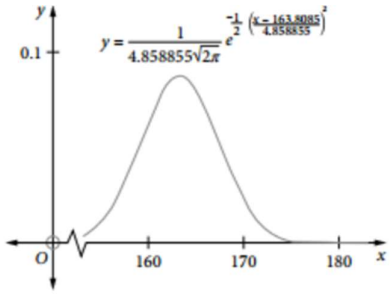


The area under the curve is given by:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-163.8085}{4.858855}\right)^2} dx = \int_{133}^{193} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-163.8085}{4.858855}\right)^2} dx$$

$$= 4.858855$$

The area should be related to the total probability (1), but it is too big by a factor of the standard deviation. Divide by this to obtain a vertical dilation which reduces the area to one, so that $f(x) = \frac{1}{4.858855\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-163.8085}{4.858855}\right)^2}$ is the final result. This is a pdf with a mean of 163.8085 and a standard deviation of 4.858855.



You would not want to go through this long process every time you have a data set to analyse. However, you can generalise the rule for this function to the form: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

The distribution represented by the pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ over domain $\mathbb{R}$, where μ is the mean of the distribution and σ is the standard deviation of the distribution, is known as the **normal distribution**. The special case where $\mu = 0$ and $\mu = 1$ is called the standard normal distribution, as all other normal distributions can be obtained from it by translation and dilation.

Note that this distribution is defined over $\mathbb{R}$, that is from −∞, to ∞. In practical situations which can be modelled by a normal distribution, there will usually be an interval $[a, b]$ for which the value of the pdf can 'reasonably' be interpreted in that situation.

The normal distribution is the most useful of all the probability distributions for continuous random variables. As has been seen, its graph is characterised by a symmetrical 'bell' shape. This shape can be used as a model for the data collected from many naturally occurring variables, such as the height of a population, the intelligence quotient (IQ) of a population, or the distribution of errors in a production process. It is also frequently used as an approximation for the sums and averages of samples taken from fixed populations and can be essential when making inferences about populations from samples.

THE NORMAL DISTRIBUTION

The symmetrical nature of the graph means that the mean and median coincide for the normal distribution and this value becomes the axis of symmetry. It should also be noted that, this graph extends infinitely (in theory) in both positive and negative directions, but it is always located above the x-axis. Thus, the normally distributed variable can assume any value.

Looking at the graph above it can be seen that the maximum value occurs when $x = \mu$. If you substitute this into the equation you get $f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}e^0$.

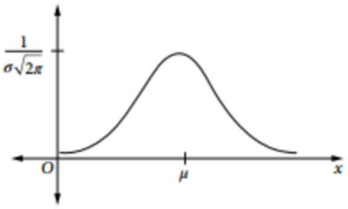
This simplifies to $\frac{1}{\sigma\sqrt{2\pi}}$, as shown in the diagram on the next page.

(This maximum value is useful when you are setting the window size for your graphing software.)

You need to check that the conditions for a pdf have been met.

By inspection, $f(x) \geq 0$ for all values of x and $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 1$.

However, the proof of this is beyond the scope of this course.



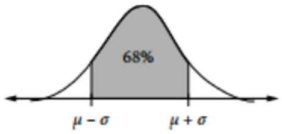
These investigations show that:

- increasing μ shifts the graph to the right, i.e. μ affects the location of the curve
- increasing σ makes the curve flatter and wider, but does not alter its axis of symmetry. If the size of the standard deviation is increased then the range $\mu \pm 3\sigma$ also increases, thus demonstrating the increased width required for the graph.

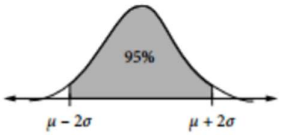
The common feature for all of the graphs is that the area under the curve is always 1.

Remember that the total area under the curve is 1 (a condition of being a pdf). Probability is related to the proportion of the total area which is being considered in a particular example.

One feature of the normal distribution that is frequently used, especially in situations where technology cannot be used, is related to the percentage of observations that you would expect to be within a certain number of standard deviations of the mean.



Approximately 68% of the observations will lie within one standard deviation of the mean.



Approximately 95% of the observations will lie within two standard deviations of the mean. It can be said that it is *very likely* that any particular observation will lie within this range.

Approximately 99.7% of the observations will lie within three standard deviations of the mean. It can be said that it is *almost certain* that any particular observation will lie within this range.

For instance, if you know that the height of a particular population is normally distributed with a mean of 170 cm and a standard deviation of 6 cm, you would expect observations roughly as follows:

- 68% in the range $\mu \pm \sigma$, i.e. 170 ± 6, which gives the range 164 to 176 cm
- 95% in the range $\mu \pm 2\sigma$, i.e. 170 ± 12, which gives the range 158 to 182 cm
- 99.7% in the range $\mu \pm 3\sigma$, i.e. 170 ± 18, which gives the range 152 to 188 cm.

You can use the 68%, 95% and 99.7% rules to help estimate and check values in many situations.

There is a short way of writing that a random variable has a normal distribution:

A continuous random variable that has a normal distribution with mean μ and variance σ^2 is written as: $X \sim N(\mu, \sigma^2)$.

Example 6

The time taken in seconds, X, for all competitors to finish a 100 m race at the school athletics carnival was found to follow a normal distribution where $X \sim N(15, 4)$. Find the following values.

(a) The time range in which you would expect to find the middle 68%.

(b) The percentage of students you would expect to take more than 19 seconds.

Solution

(a) Write the mean μ and the standard deviation σ: $\mu = 15$, $\sigma^2 = 4$, $\sigma = \sqrt{4} = 2$
The middle 68% describes the values within one standard deviation of the mean, i.e. $\mu \pm \sigma$:
15 ± 2 gives the range 13 to 17 seconds.
You would expect the middle 68% of participants to take between 13 and 17 seconds to complete the race.

(b) Identify the value in terms of the mean and the standard deviation: $19 = \mu + 2\sigma$
(You may need to use a guess-and-check process.)
Use the symmetry of the curve to answer the question.
(In this case you are using only one tail, so you need to halve the percentage.)
You would expect 5% outside the range $\mu \pm 2\sigma$, so you would expect 2.5% of competitors to take more than 19 seconds.

Beware:

The formal description of the normal distribution $N(\mu, \sigma^2)$ uses the variance σ^2 as a parameter. However, worded questions often give the standard deviation σ. Make sure you read the questions carefully. When using technology you will also usually need to use the standard deviation and not the variance.