

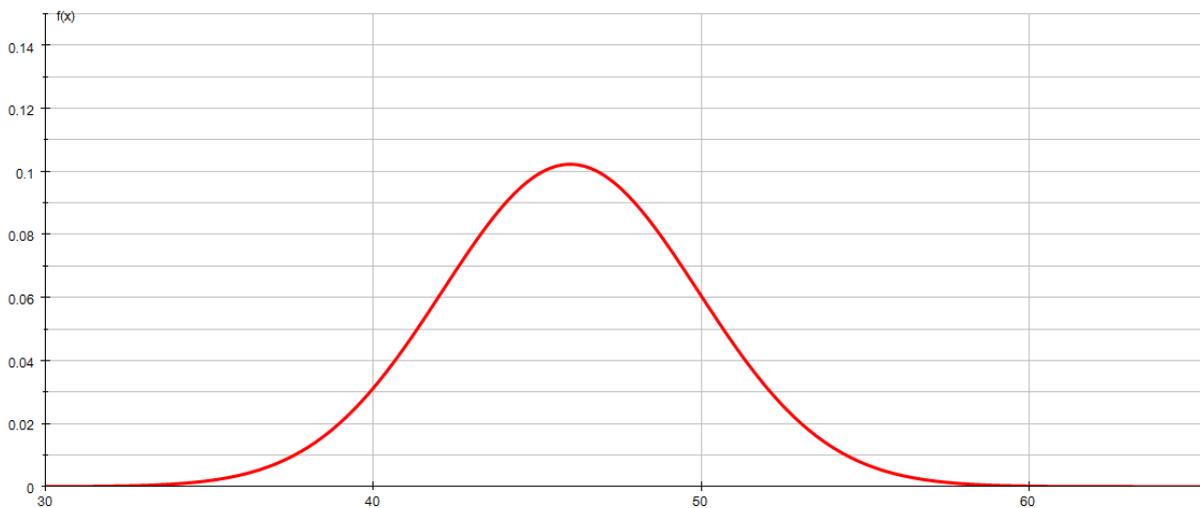
The Further Mathematics Support Programme

Fly Lengths

Standard Deviation and the Normal Distribution



FREE IMAGES UK



A lot of biological data follows a **Normal Distribution** which has the characteristic bell shaped curve above. Typically heights and masses of different species will follow a distribution like this, although one has to be careful to ensure that other variables such as age and gender are taken into account, so we might consider the heights of women aged 40 to 45 for example. Another example might be the widths of leaves of a particular species of tree.

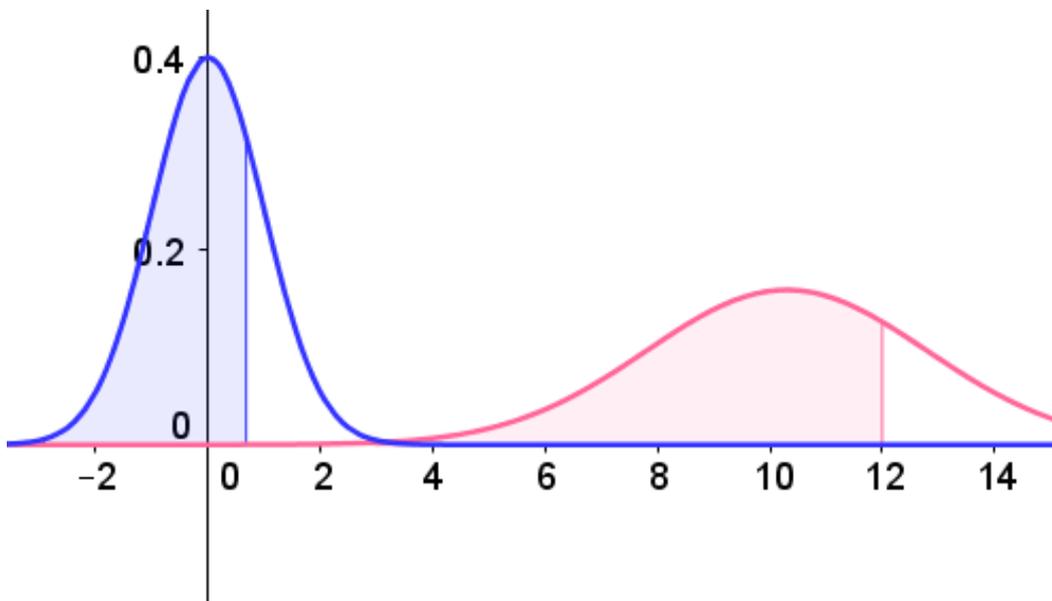
The normal distribution has a number of interesting properties including:

- It is symmetrical about the mean
- The mean, mode and mean are all the same

Usually it would be impossible to access a complete **population**, so most scientists would take a **sample** of data and try to use that to decide whether it is reasonable to assume that the population as a whole follows this distribution. Establishing the distribution of the population allows lots of other conclusions to be drawn from the data and predictions to be made, often with relatively small amounts of data.

There are some quite sophisticated ways of testing data to see how well it fits this shape or distribution but there are also more simple checks that can be performed initially to get a general idea before more advanced work is undertaken. This check uses some of the remarkable properties of the normal curve and involves calculating the mean of the data and how spread out the data is from the mean, which can be measured using the standard deviation.

Different data sets may give rise to different normal distributions:



In this example the means are different and the data is spread out differently about the mean. In the case of the right hand curve the data is much more widely spread out about the mean of 10.3, which is in the middle of the distribution.

In order to be more specific about the amount of 'spread' we need to have a quantity which measures it, just like we have the mean that measures the central point or 'location' of the data.

At GCSE you will have worked out the range and interquartile range of the data and have used these to say how spread out the data is. These are quite useful measures of spread but they only use some of the data items, not all of them and we tend to use these with the median.

A better measure of spread, which like the mean, uses every data item is the **standard deviation**.

This is given by the formula:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Where \bar{x} is the mean of the data and n the number of data items. The standard deviation is s and s^2 is called the variance.

There are different versions of this formula so it may appear slightly differently in some text books.

It is probably best to look at some data to see how we calculate this. The data below has been collected from a sample of 100 Houseflies and gives their wing lengths in mm:

| | | | | | | | | | | | | | | | | | | | | |
|------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Length(mm) | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 |
| Frequency | 1 | 1 | 2 | 3 | 4 | 6 | 7 | 8 | 10 | 10 | 11 | 9 | 8 | 7 | 6 | 3 | 2 | 1 | 0 | 1 |

Using a spreadsheet we can calculate that $\sum x = 4521$ and so $\bar{x} = 45.21$

Again using a spreadsheet we can then work out that $\sum(x - \bar{x})^2 = 1398.59$

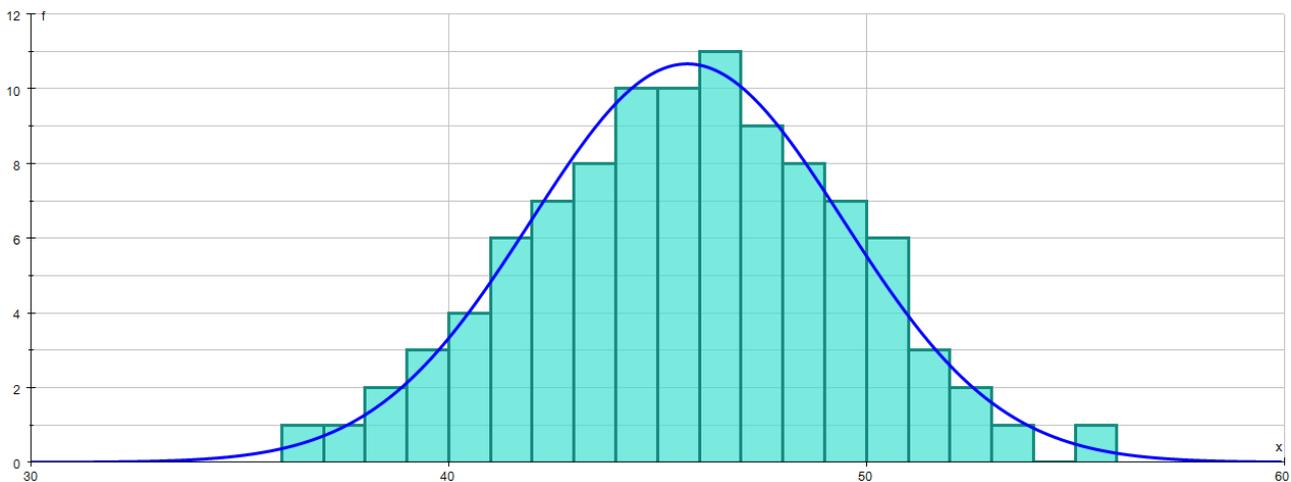
| Length(mm) | Frequency | | |
|------------|------------|--------------|--------------------|
| x | f | xf | (x-m)^2*f |
| 36 | 1 | 36 | 84.8241 |
| 37 | 1 | 37 | 67.4041 |
| 38 | 2 | 76 | 103.9682 |
| 39 | 3 | 117 | 115.6923 |
| 40 | 4 | 160 | 108.5764 |
| 41 | 6 | 246 | 106.3446 |
| 42 | 7 | 294 | 72.1287 |
| 43 | 8 | 344 | 39.0728 |
| 44 | 10 | 440 | 14.641 |
| 45 | 10 | 450 | 0.441 |
| 46 | 11 | 506 | 6.8651 |
| 47 | 9 | 423 | 28.8369 |
| 48 | 8 | 384 | 62.2728 |
| 49 | 7 | 343 | 100.5487 |
| 50 | 6 | 300 | 137.6646 |
| 51 | 3 | 153 | 100.5723 |
| 52 | 2 | 104 | 92.2082 |
| 53 | 1 | 53 | 60.6841 |
| 54 | 0 | 0 | 0 |
| 55 | 1 | 55 | 95.8441 |
| | 100 | 4521 | 1398.59 |
| | n | sum x | sum (x-m)^2 |
| | | | 45.21 |
| | | | mean m |

Hence $s = \sqrt{\frac{1398.59}{99}} = 3.76$

Here we have calculated \bar{x} the mean of the sample and s the standard deviation of the sample. We don't actually know the mean μ and standard deviation σ of the population from which this data is drawn but we use \bar{x} and s as best estimates of μ and σ respectively. These are known as **unbiased estimators**.

So are the fly lengths normally distributed?

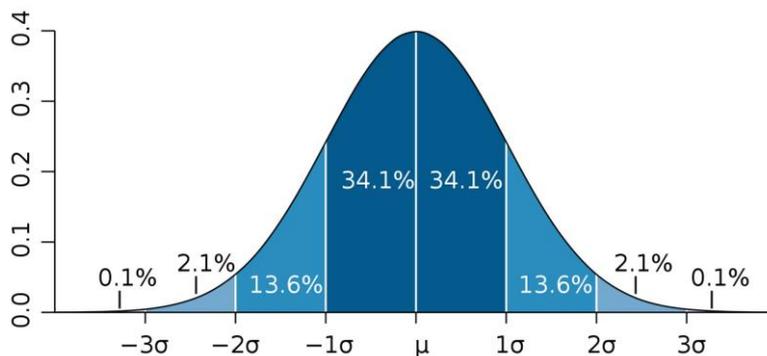
We could firstly plot the lengths and look at the shape of the resulting graph:



This looks a good fit to a normal curve.

There are other properties of a normal distribution that we could use:

- 68% of the data lies within one standard deviation of the mean
- 95% of the data lies within two standard deviations of the mean
- 99.7% of the data lie within three standard deviations of the mean



In theory the curve extends infinitely in both directions, but the chance of finding data items outside the 3 standard deviation mark is very small.

So let's use the standard deviation to check some of these properties:

Number of data items within 1 S.D. of the mean:

$\bar{x} \pm s$ gives $45.21 - 3.76$, $45.21 + 3.76$) as an interval which is (41.45,48.97)

There are roughly 66 data items in this interval.

Here we have taken half the items in the 41mm class together with 42mm to 48mm.

There is a slight discrepancy here as the lengths have been measured to the nearest mm and the normal distribution is continuous. Nevertheless 66% is very close.

Number of items with 2 S.D. of the mean:

$\bar{x} \pm 2s$ gives ($45.21 - 7.52$, $45.21 + 7.52$) which is (37.69,52.73)

There are roughly 96 items in this in interval which agrees with the expected 95% very closely.

You could also check 3 S.D. if you wanted but is probably clear that the interval would contain all the data.

So we can conclude from the shape and approximate symmetry of the graph, the fact that it has one peak (or mode) and from the proportions lying with 1, 2 and 3 S.D of the mean that it follows a normal distribution.

References

Seattle Central Edu "Quantitative Environmental Learning Project."

<http://www.seattlecentral.edu/qelp/sets/057/057.html> [Accessed: 30/1/15]

Sokal, R.R. and F.J. Rohlf, 1968. Biometry, Freeman Publishing Co., p 109.

Sokal, R.R. and P.E. Hunter. 1955. A morphometric analysis of DDT-resistant and non-resistant housefly strains Ann. Entomol. Soc. Amer. 48: 499-507.