

MBBS Stage I: notes on the Normal Distribution

Samples and Populations

In the first row of figure 1 the percentage cumulative frequency curve and a percentage relative frequency histogram of a sample of 63 haematocrit values taken from male stage I medical students are shown. These give a reasonable description of the distribution of this clinical variable in a population of young adult males: the distribution is centred, or located, at about 46% to 47% and most of the values are between about 44% and the 51%. A more precise description is that the median is 46% and the lower and upper quartiles are 45% and 48% respectively. Although these values are useful as some indication of the variation in the given population, they are based on a single sample, and values obtained from another sample, perhaps from a different year or different medical school, may well be different. The sample is thought of as providing an *estimate* of the underlying population of young adult males: graphs in row 1 of figure 1 provide estimates the underlying population versions of the curves, shown in row 2.

The vertical axis of the population cumulative curve shows the percentage of the population whose haematocrit value falls below the corresponding point on the horizontal axis. The population relative frequency curve is perhaps more intuitive, as it clearly conveys the impression of most values falling near the peak of the curve, and progressively fewer as the values move away from the centre. It is the natural population analogue of the sample relative frequency histogram. Its precise definition is surprisingly complicated but the loose description just given is sufficient for the present.

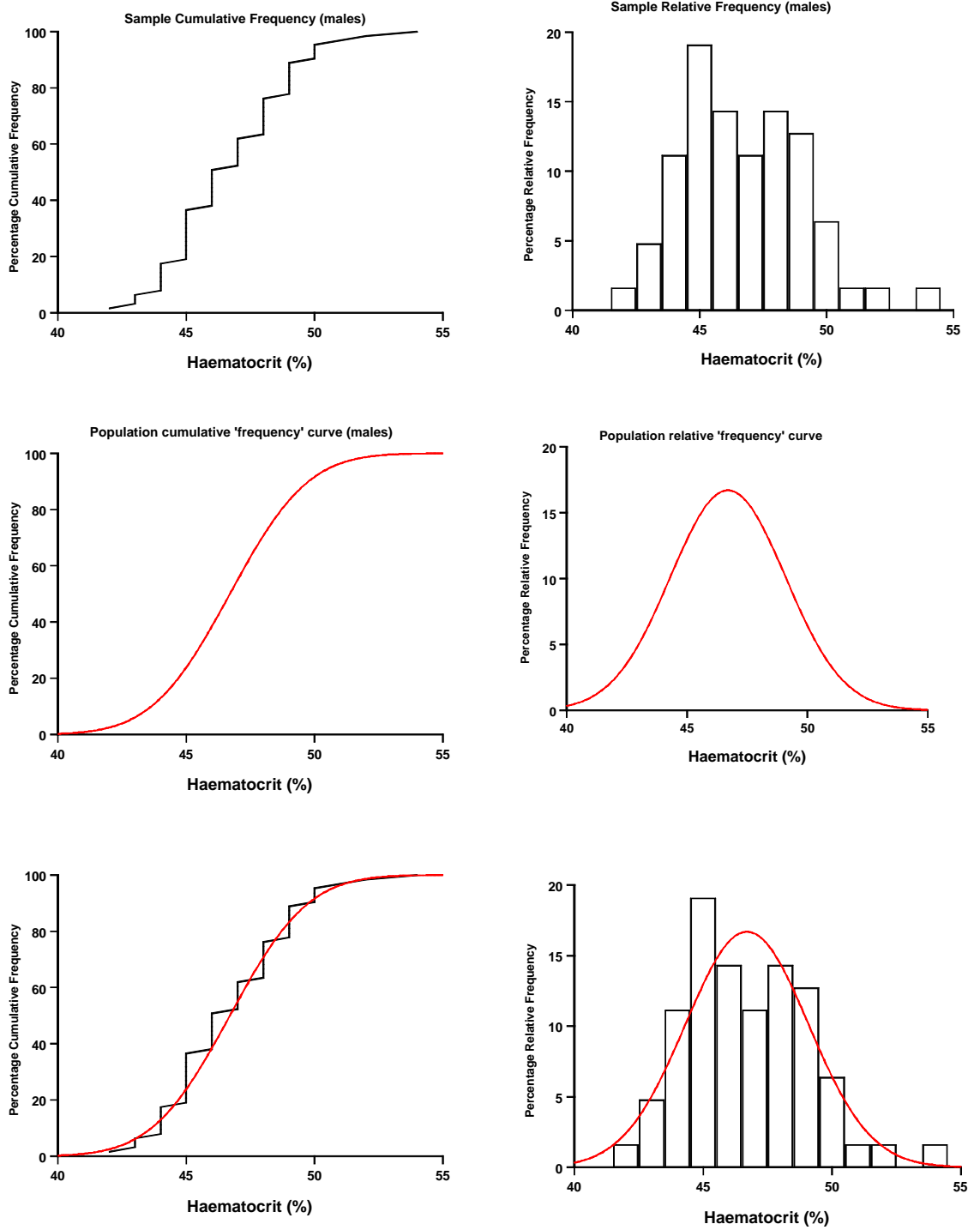
Population curves are never known exactly and those shown in figure 1 are hypothetical. Either of the two curves in row 2 defines the *distribution* of values in a population and many shapes of distribution are possible. Those shown correspond to a particular distribution, known as the Normal distribution; it is very commonly used and one of many reasons for this will be outlined below. The Normal distribution is sometimes called the Gaussian distribution but the former term will be used here, with the capital letter to show that in this context the word 'normal' has now acquired a technical meaning.

The Normal Distribution

What does it mean to say that a variable, e.g. haematocrit, follows a Normal distribution? Roughly speaking it means that most values in the population are close to that of the single central peak and values get steadily less common as they move away from the centre. Values the same distance either side of the peak are equally common, i.e. the distribution is *symmetric*. Not all distributions are like this, and two alternatives are shown in figure 2: the one on the left is called a *skew* distribution and the one on the right is a *bi-modal* distribution. Skew distributions are encountered quite often in medicine, for things such as skin-fold measurements and bilirubin values: bi-modal and other distributions occur occasionally. However, many common medical variables, such as heights, haemoglobin concentrations, haematocrits and variables from clinical chemistry have a symmetric distribution about a single central peak, that is a Normal distribution*.

* Symmetric, single-peak distributions exist that are not Normal, but for practical purposes of data-description these can be ignored.

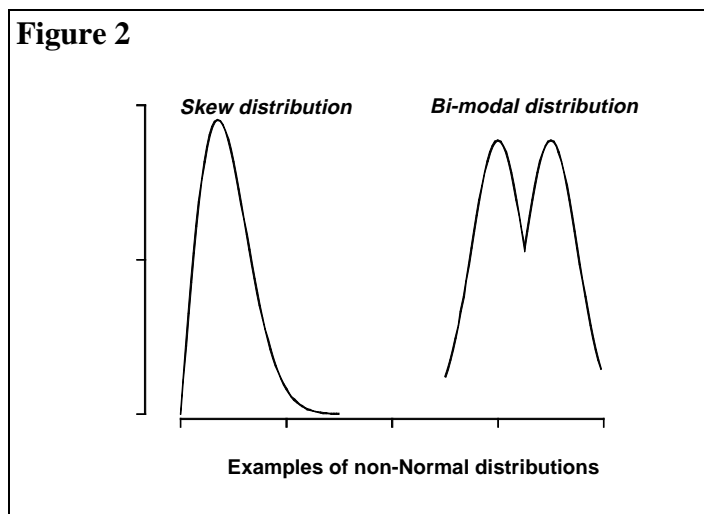
Figure 1



Male haematocrit values: 'cumulative' representation on left. Sample cumulative percentage frequency and percentage relative frequency histogram. The smooth curves are possible population analogues of the sample curves.

One point that should be made is that, strictly speaking, only variables that can take 'any' value, such as height or haemoglobin concentration, can possibly have a Normal distribution; these are referred to as *continuous variables*. Variables such as blood group or eye colour which can take only a few distinct values, so-called *discrete variables*, cannot have a Normal distribution. In practice Normal distributions are often applied to variables, such as haematocrit, which are in principle continuous (in theory they can take any value from 0 to 100%) but which can be measured with only limited accuracy, so giving only whole-number percentage values.

There is no single reason why so many biological variables have a Normal distribution. One reason is connected with the genetic control of continuously varying attributes, such as height and this is explained in more detail in the next section. Another is that measurements are



often the sum of many smaller components, e.g. the haematocrit measurement is the sum of the volumes of the packed red cells. This form of aggregation leads to Normal distributions, although why this is so is related to deeper properties of the Normal distribution that are beyond the scope of this note. Another reason is simply observation, i.e. the shapes of distributions of many commonly measured quantities have, over many years, been observed to conform to the pattern seen in row 2 of figure 1.

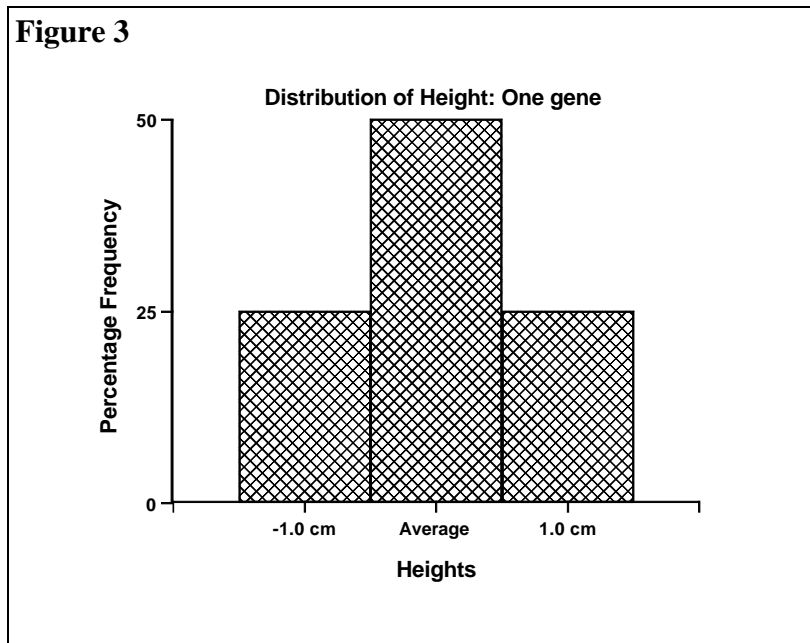
A Genetic basis for the Normal Distribution.

This section presents an explanation of the way in which some types of genetic control of continuously varying attributes can lead to distributions that appear Normal; height is taken as the example.

The variability of some discrete variables, such as Rhesus blood groups, $Rh+$ or $Rh-$, are controlled by the action of a single gene. There are alleles D and d , with D dominant; $Rh+$ results from DD and Dd , with dd giving $Rh-$. In this example the heterozygous form is phenotypically indistinguishable from the dominant homozygote. However, it is possible for an attribute under the control of a single gene to exhibit three phenotypes, that is the heterozygote is distinguishable from both forms of homozygote (a clinically important example is sickle-cell anaemia[§]).

[§] For details, see Fraser Roberts and Pembrey, *An Introduction to Medical Genetics*, Oxford, chapter 3

For illustrative purposes, suppose for the moment that the inheritance of height is under the control of a single gene with alleles H and h . Suppose also that individuals with genotype Hh are phenotypically of average height, that a genotype HH results in a phenotype 1cm taller than average and hh in a phenotype 1cm shorter than average. There would then be only three heights in the population, namely average (Hh), 1 cm below average (hh) and 1 cm above average (HH). If the alleles H and h are equally prevalent each combination HH , hh , Hh and hH is also equally



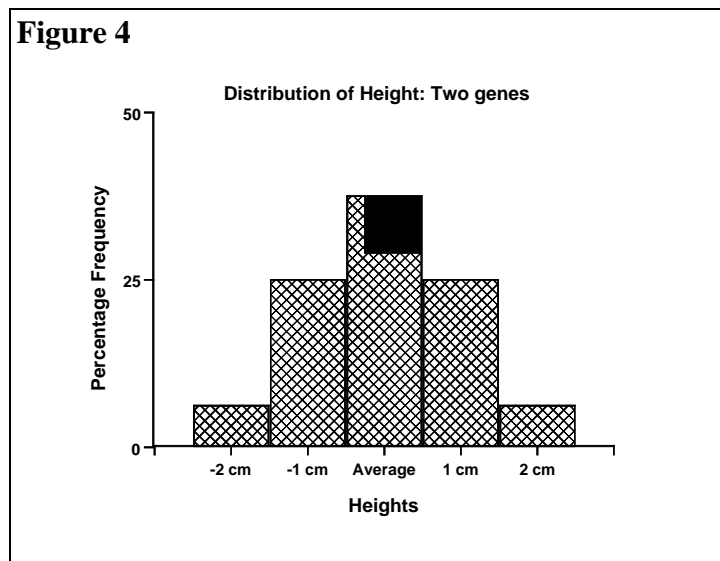
likely (where hH and Hh have been used to distinguish the heterozygote where h comes from, respectively the mother or father). However, Hh and hH both have average height, so the final distribution of the phenotypes is as in figure 3.

Suppose now that instead of just one gene controlling height, two are needed, again each

		Gene 1			
		hh	hH	Hh	HH
Gene 2	hh	-2cm	-1cm	-1cm	0cm
	hH	-1cm	0cm	0cm	1cm
	Hh	-1cm	0cm	0cm	1cm
	HH	0cm	1cm	1cm	2cm

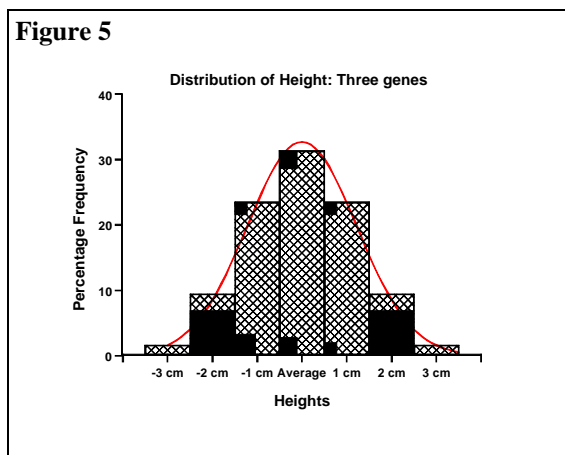
with alleles H or h . The height of the phenotype is determined by the excess of the number of H alleles over the number of h alleles: equal numbers lead to average height, two more H than h results in an individual 1 cm above average, two more h than H results in an individual 1 cm below average, four more h than H gives a phenotype 2 cm below average and so on. The possible outcomes are given in the table below: the entries in the body of the table are the departures from average height (so 0cm = average) of the phenotype corresponding to the genotypes obtained from the forms of genes 1 and 2 along the margins of the table. Each of the $4 \times 4 = 16$ possible combinations of gene 1 and 2 is equally likely, but these give rise to only five different heights, namely average and 1 and 2 cm above and below the average. As only one of

the sixteen possible outcomes gives an individual 2 cm above average, we know that only $1/16 \times 100\% = 6.25\%$ of the population are of this height, whereas 6 of the outcomes, or



$6/16 \times 100\% = 37.5\%$, have average height. The full distribution is shown in figure 4.

If the number of genes controlling height is now supposed to be 3, there are $4 \times 4 \times 4 = 64$ equally likely gene combinations, but these give rise to only seven phenotypes, namely heights at 1cm intervals from -3cm to 3 cm. By counting the number of gene combinations giving rise to each height, we can construct the height distribution for this population, as we did above for one and two gene control of height above. The distribution for three genes shown in figure 5 is beginning to look quite like a Normal distribution, as the superimposed Normal curve indicates.

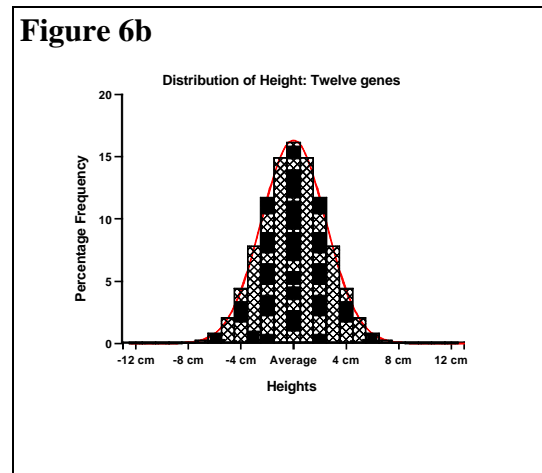
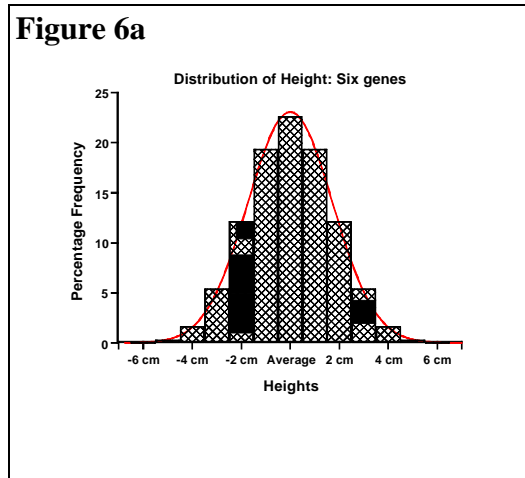


It is possible to extend this argument to any number of genes controlling height and figure 6 a) and b) show the distributions obtained when respectively 6 and 12 genes control height. Clearly, as the number of genes controlling height increases, the number of possible heights increases

and their distribution gets closer and closer to a Normal distribution. This is an example of the *polygenic* control of a continuously varying attribute.

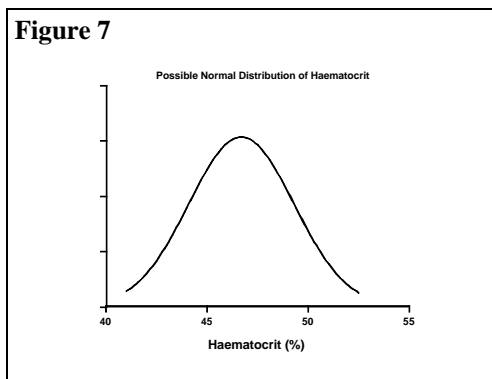
Of course, this is a greatly simplified model of how height is inherited because many important aspects have been ignored, including aspects of the influence of parental height on that of the offspring and the assumption that each gene contributes the same amount to the final height. Perhaps even more important is that the final height of an individual is not wholly determined by genetic factors but is also influenced by environmental factors, such as nutrition and healthcare. It should also be realised that if an attribute, such as height, has a Normal distribution it does not follow that it is under polygenic control, nor if an attribute has, e.g. a skew distribution, does it mean that the attribute is not genetically influenced to some extent.

Nevertheless, the preceding argument shows that a Normal distribution can occur as the result of a biologically plausible mechanism.

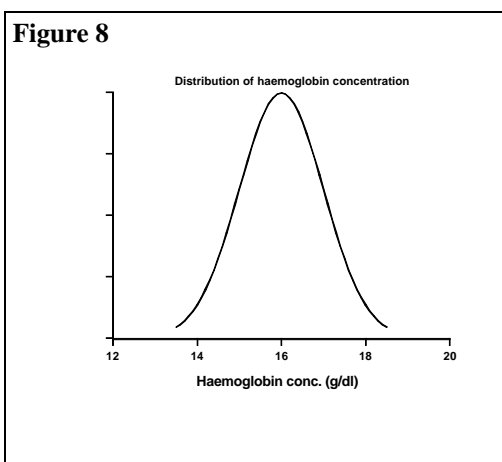


Different Normal Distributions

Although it is difficult to tell from samples of the size analysed in figure 1, it is actually reasonable to assume that the distribution of male haematocrit values in a population of young adult males in the UK follows a Normal distribution. A possible version is shown in figure 7. It is

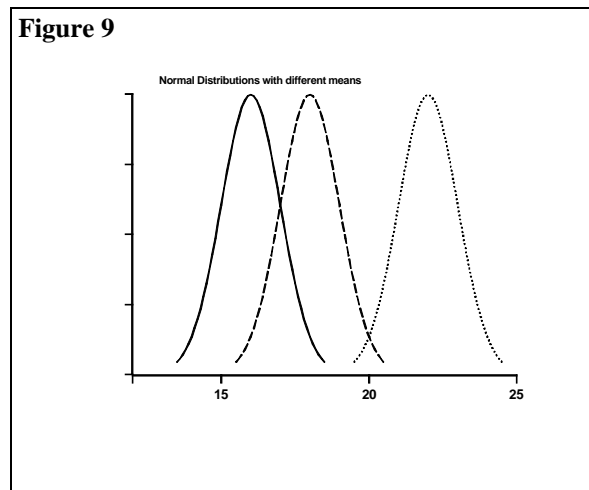


also reasonable to assume that haemoglobin concentration in this population follows a Normal distribution and this appears in figure 8. However, it is clear that they are different distributions, the former describes a variable whose values are located around the 40s, whereas the latter is around the mid teens. Because these are both Normal the two distributions share the same general features and are, in a fairly obvious sense, the same 'shape'. So what is it that makes the two distributions different? There are essentially two things: i) the position of the central peak and ii) how dispersed the data are about this central value. That is, the location and spread of the *population*.

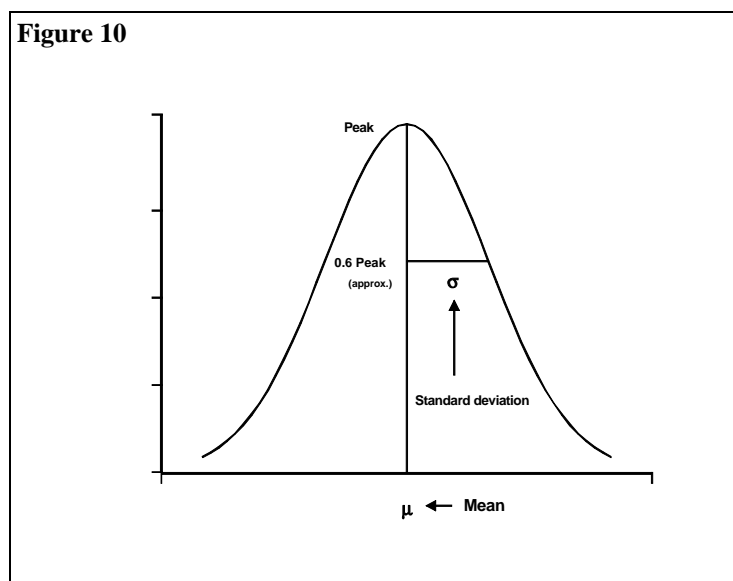


The value at the central peak is called the *population mean* and is often denoted by the Greek letter μ (mu). The value of this quantity is unknown as it refers to the entire population. It is an example of a *population parameter*; it is conventional to use Greek letters to refer to unknown population parameters. If a sample of data from the population is available then we can *estimate* μ - how this is done is dealt with in the last section of this note. Figure 9 shows several Normal distributions that all have the same dispersion but different means.

While members of the population will have values that vary either side of μ , how *far* they stray from the central value is determined by the other population parameter of the Normal



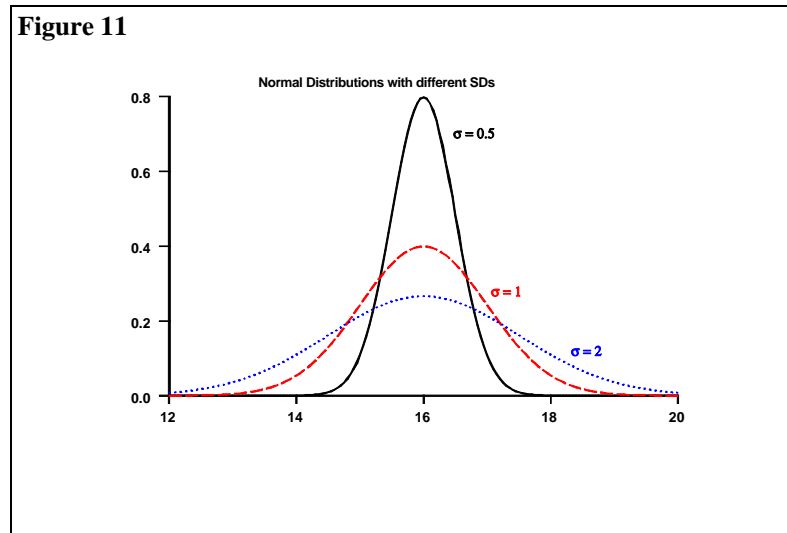
distribution, which is usually written as σ , the population *standard deviation* (SD). How an estimate of σ can be found from a sample is given in the final section, but for the present it should simply be thought of as a measure of spread. This is obviously some measure of the width of the bell-shape of the distribution. It does not really matter exactly where or how the width is measured, provided it is defined in a way that is consistent and can be reproduced. In figure 10, the definition of σ is demonstrated: it is half the horizontal width of the 'bell-shape' at a point that is at 60.6%* of the height of the peak. The reason for this rather strange definition is that it simplifies the formula that is used for the estimation of σ from a sample.



In figure 11 the effect of changing σ on the appearance of the distribution is illustrated. Note that because σ measures a width it cannot be negative, unlike μ which can take any value.

Thus the shape of the distribution is defined by the fact it is Normal, its location by μ and its spread by σ . Once these three things are known, the distribution is completely specified. In particular all the centiles of the distribution are known. For example, a feature of the Normal distribution is that 16% of the population falls below the point that is one SD below the mean, i.e. $\mu - \sigma$. If a point even further away from the mean is chosen, namely two SDs below the mean,

* the value 60.6% is an approximation to $\exp(-1/2)$



$\mu - 2\sigma$, then 2.3% of the population lies below this point. Complementary statements can be made about points above the mean from the symmetry of the Normal distribution, see figure 12a. This figure shows that about 2.3% of the population is more than two SDs below the mean and that about 16% are less than one SD below the mean. What point cuts off the bottom 10% of the population? Clearly, it will be between one and two SDs below the mean, that is it can be written as $\mu - z\sigma$ where z is a value between 1 and 2. In fact the required value is 1.28. The values of z needed to cut off various percentages, such as 3%, 10%, 25% are shown in figure 12b.

Figure 12a

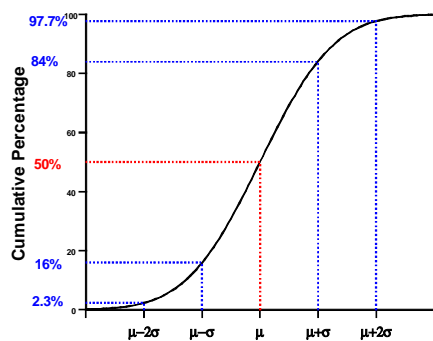
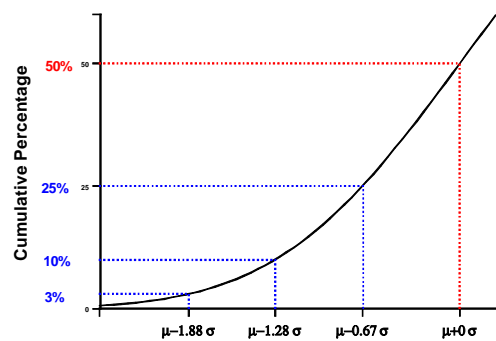


Figure 12b



The important point to note about the above is that it is quite general, the results are true for any μ and σ , and hence for any Normal distribution. The proportion of the population that is less than $\mu + z\sigma$ depends only on the value of z : if z is 0, the proportion is obviously 50% (the Normal curve is symmetric about its peak), a negative value of z corresponds to percentages less than 50 and a positive value to a percentage exceeding 50. Some values are given in figure 12,

z	0	-1.96	-2.32	-2.58	1.96	2.32	2.58
Percentage below $\mu + z\sigma$	50%	2.5%	1%	0.5%	97.5%	99%	99.5%

some others are shown in the table below. Unfortunately, there is no simple formula that relates the value of z the proportion of the population that is cut off by $\mu + z\sigma$; it is necessary to use

either a computer or a set of tables. In fact, most of the commonly used values appear in figure 12 or the above table, or can be deduced from values found there.

How can we use these results?

Suppose it is required to find the 3rd centile of the distribution of haematocrit levels of young adult males in UK. If haematocrits follow a Normal distribution then it will be $\mu - 1.88\sigma$. The problem is that because μ and σ refer to the population, neither is known. If a sample of values from the population is available then the way round this problem is to use the sample to find an estimate of the 3rd centile. One possibility is to use the sample cumulative frequency curve, as in figure 1, but to estimate such an extreme centile this way clearly needs a very large sample. An alternative is that if we *assume* that haematocrit values are Normally distributed, then there is an alternative, namely that the sample is used to produce estimates m and s of μ and σ respectively, and the 3rd centile is then estimated as $m - 1.88s$. The case $z = 1.96$ is important as $\mu \pm 1.96\sigma$ demarcates the central 95% of the population and is often used by clinical chemists, and others to define reference ranges

Estimating μ and σ from a sample

As μ is the population mean, it is natural to estimate it by the mean of the sample, i.e. if the sample comprises values x_1, x_2, \dots, x_n , then the sample mean m is simply

$$m = \frac{x_1 + x_2 + \dots + x_n}{n}$$

How σ is estimated from the sample is a little less obvious. Clearly we want to measure how far the individual haematocrit values depart from the centre of the distribution, that is from μ , so ideally the method should be based on getting a 'typical' or average value for the distances from the central value, i.e. the average of $x_1 - \mu, x_2 - \mu, \dots, x_n - \mu$. Such distances cannot be computed because μ is unknown, but we now have an estimate of it, namely m . So why not find the average of $x_1 - m, x_2 - m, \dots, x_n - m$? The snag is that the sample average of these values is always 0 (try it on a sample of any three numbers); the problem being that values above m have a positive departure, whereas those below m have a negative departure and these cancel out exactly.

The way σ is estimated is to average $x_1 - m, x_2 - m, \dots, x_n - m$ in some way but without taking account of the sign of the departure. It would be possible simply to average the values after knocking the sign off the negative values. This leads to technical problems, so what is done is to square each of the departures (which is another way to make them all positive), and to average these quantities. There are two further complications. First, the average of the squared values is in squared units (e.g. if the sample were weights, $(x_i - m)^2$ is in kg^2) and this is not what we want for a measure of spread, so the square root of this average must be taken. Second, for a sample of size n , we form the average by dividing not by n but $n-1$. The reason for this is rather obscure and its full explanation is beyond the scope of this note; it is to do with the need to use m rather than μ in the calculations.

To summarise this, we estimate σ by s where

$$s^2 = \frac{(x_1 - m)^2 + (x_2 - m)^2 + \dots + (x_n - m)^2}{n - 1}$$

An example

The following is a sample of 15 haematocrit values from young adult males. The sample mean and the sample standard deviation are found as:

46	51	49	48	50	46	45	45
49	43	44	45	46	48	48	

Mean

$$m = \frac{46 + 51 + 49 + \dots + 48}{15} = 46.9$$

Standard Deviation

$$\begin{aligned} s^2 &= \frac{(46 - 46.9)^2 + (51 - 46.9)^2 + \dots + (48 - 46.9)^2}{15 - 1} \\ &= \frac{0.9^2 + 4.1^2 + \dots + 1.1^2}{14} \\ &= 5.41 \\ \therefore s &= \sqrt{5.41} = 2.3 \end{aligned}$$

While it can be useful to perform these calculations by hand when trying to understand the definitions, it is not necessary to do this routinely: programs such as Minitab and Excel can compute them more easily *and more reliably*.

A consequence is that the 3rd centile of the distribution can now be estimated as:

$$m - 1.88s = 46.9 - 1.88 \times 2.3 = 42.6.$$

Any other centile can be calculated in an analogous way, once the appropriate z value (known as the *Standard Normal Deviate*) has been found from tables similar to, but possibly more extensive than, the one above.

Summary

The Normal distribution is a particular form of distribution that is found in many branches of biology and medicine. The particular Normal distribution is specified once its two parameters, the mean and standard deviation, are known. Estimates of these may be computed easily from a sample of values from the population. *If* a variable comes from a Normal distribution then it is possible to estimate any of its centiles using appropriate combination of the sample mean and standard deviation.

The advantage of assuming a Normal distribution is that more of the information in the sample is used when making inferences about all aspects of the distribution, including e.g. extreme centiles. Using a cumulative frequency graph to estimate such quantities will often rely heavily on the values of the few smallest values in the sample, whereas the estimate $m - 1.88s$ depends on all the sample, through m and s . The precision of such estimates is much greater than those obtained from a percentage cumulative frequency curve.

The disadvantage is that all the benefits stem from making the assumption of Normality, and it can be very difficult to check if this is reasonable. Plotting histograms is one way; more

sensitive methods exist but are beyond the scope of this note. The reason that the Normal distribution is so important is that the assumption is often justified.

Further Reading

Armitage, P. and Berry, G. (1994) *Statistical Methods in Medical Research*, Blackwell, p.67-
Bland, Martin (1987) *An Introduction to Medical Statistics*, Oxford, chap. 7.