

Data Description, Populations and the Normal Distribution

Introduction

This course is about how to analyse data. It is often stressed that it may be totally impossible to produce a meaningful analysis of a set of data, or at least it may not be possible to use the data to answer questions of interest, unless the data have been collected properly. This is undoubtedly so, and correct design of studies is fundamentally important. However, although correct design must obviously precede correct analysis in the conduct of any investigation, the principles of design are there simply to facilitate correct and efficient analysis, so a sound understanding of how data are analysed is needed in order to appreciate the tenets of sound design. It is for this reason that this course concentrates on issues of analysis.

Descriptive Statistics: the five number summary

Once a set of data has been collected, one of the first tasks is to describe the data. The following table contains the heights of 99 five-year-old British boys in cm.

| | | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 117.9 | 110.2 | 112.9 | 115.9 | 108.0 | 104.6 | 107.1 | 117.9 | 111.8 |
| 106.3 | 111.0 | 100.4 | 112.1 | 109.2 | 101.0 | 105.4 | 99.4 | 110.1 |
| 103.3 | 106.9 | 108.2 | 119.3 | 112.0 | 106.2 | 105.9 | 106.9 | 109.3 |
| 105.9 | 110.0 | 106.7 | 108.5 | 107.7 | 114.3 | 108.6 | 104.6 | 113.7 |
| 116.7 | 103.5 | 96.1 | 110.8 | 97.2 | 109.6 | 110.5 | 105.9 | 106.2 |
| 107.4 | 114.9 | 110.3 | 104.8 | 99.2 | 119.2 | 111.4 | 103.0 | 110.1 |
| 105.8 | 101.5 | 105.9 | 107.6 | 97.1 | 113.3 | 109.4 | 109.4 | 110.8 |
| 106.3 | 108.1 | 109.6 | 102.4 | 110.4 | 110.1 | 115.3 | 102.9 | 111.2 |
| 99.4 | 105.7 | 119.5 | 109.3 | 112.8 | 108.2 | 117.0 | 106.8 | 105.4 |
| 108.7 | 109.2 | 97.1 | 103.3 | 108.8 | 116.3 | 115.5 | 114.9 | 101.1 |
| 104.1 | 110.8 | 112.7 | 105.6 | 99.9 | 111.1 | 109.4 | 109.1 | 110.7 |

The immediate impression is of an indigestible mass of numbers. Some of the numbers are under 100 cm, a few are above 115 cm but most seem to be between 100 cm and 115 cm. Little more can be said from this display. Some progress can be made by re-arranging the table so that the heights are in numerical order, as in the following:

| | | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 96.1 | 101.5 | 105.4 | 106.3 | 108.1 | 109.3 | 110.3 | 111.8 | 115.3 |
| 97.1 | 102.4 | 105.4 | 106.3 | 108.2 | 109.4 | 110.4 | 112.0 | 115.5 |
| 97.1 | 102.9 | 105.6 | 106.7 | 108.2 | 109.4 | 110.5 | 112.1 | 115.9 |
| 97.2 | 103.0 | 105.7 | 106.8 | 108.5 | 109.4 | 110.7 | 112.7 | 116.3 |
| 99.2 | 103.3 | 105.8 | 106.9 | 108.6 | 109.6 | 110.8 | 112.8 | 116.7 |
| 99.4 | 103.3 | 105.9 | 106.9 | 108.7 | 109.6 | 110.8 | 112.9 | 117.0 |
| 99.4 | 103.5 | 105.9 | 107.1 | 108.8 | 110.0 | 110.8 | 113.3 | 117.9 |
| 99.9 | 104.1 | 105.9 | 107.4 | 109.1 | 110.1 | 111.0 | 113.7 | 117.9 |
| 100.4 | 104.6 | 105.9 | 107.6 | 109.2 | 110.1 | 111.1 | 114.3 | 119.2 |
| 101.0 | 104.6 | 106.2 | 107.7 | 109.2 | 110.1 | 111.2 | 114.9 | 119.3 |
| 101.1 | 104.8 | 106.2 | 108.0 | 109.3 | 110.2 | 111.4 | 114.9 | 119.5 |

This is much better and the investigator can glean a good deal of information from this presentation, such as whether there are any unusual values in the sample, as well as getting a better appreciation of the distribution of these 99 heights. However, it can hardly be said to be a succinct way to present the data, and when giving presentations, or publishing results or comparing different data sets (such as a set of heights of children from another country) and for many other purposes, more economical summaries are needed.

A common way to summarise a data set is the *five number summary*, and for these heights the five numbers are the ones highlighted in the above table.

The number in the cell with the bold double outline, 108.7 cm, is the ‘middle’ value of the sample when it is placed in ascending order: it is the 50th largest value, so 49 values are smaller than it and 49 values are larger. It is known as the *median* and is a widely used measure of the *location* of a sample.

Samples can be located similarly but be quite different because they can be more or less dispersed around their location. It is therefore useful to have a measure of the *spread* of the sample. There are several possible measures and a widely used one is provided by the quartiles, which are the numbers with the lighter double outlines in the table. The *lower quartile*, 105.6 cm, is defined as the number which is a quarter of the way from the smallest to the largest value in the sample. The *upper quartile*, 111.1 cm is three quarters of the way from the smallest to the largest value in the sample. The inter-quartile range (IQR) is defined as the difference between the figures, i.e. 5.5 cm. An alternative measure of spread, which will be seen later to have severe deficiencies, is the *range*, which is the difference between the *maximum* in the sample (119.5 cm) and the *minimum* (96.1 cm), i.e. 23.4 cm.

The exact definitions of these quantities need care so that several awkward technicalities are overcome consistently. In the present example there is a value that is unequivocally the ‘middle’ value because there are an odd number of observations in the sample. Had there been an even number, for example 100 observations, there would be problems of definition: the 50th largest number would exceed 49 values but be exceeded by 50 values, whereas the 51st largest number would exceed 50 values but be exceeded by 49 values, so neither would be exactly in the middle, but each would

have an equal claim to this status. The solution is to define the median as being half-way between the 50th and 51st values. Similar issues attend the quartiles and the requisite formulae are given in appendix 1.

The minimum, lower quartile, median, upper quartile and maximum collectively comprise the five number summary.

Graphical Displays

A useful graphical display of this way of summarising data is the *boxplot*, or *box and whisker plot*, shown in figure 1.

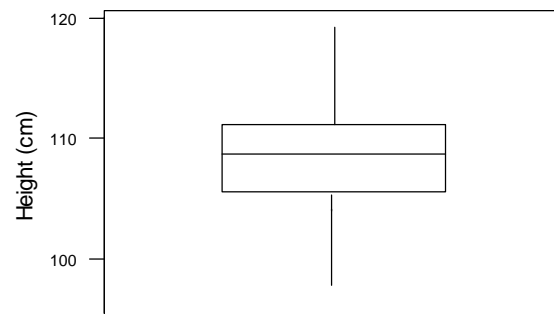


Figure 1

The top and bottom of the ‘box’ of the box and whisker plot are drawn at the level of the upper and lower quartiles respectively. A line is drawn across the box at the level of the median. This is the usual form the box and variations from this are very rarely encountered.

The ‘whiskers’ are lines drawn from the top and bottom of the box to the maximum and minimum and this is what is shown in figure 1. This is how the box and whisker plot was originally conceived. However variants on this are quite common. Usually they involve drawing the whiskers up to points that are within a given multiple of the height of the box from the top or bottom of the box. Any points beyond the whiskers are plotted individually. This approach is often adopted in computer packages and can be useful insofar as unusual or outlying values are plotted explicitly and do not simply distort the lengths of the whiskers. The exact multiple of box widths varies between packages: typical values are between one and two.

An alternative way to display data is a *histogram*. The range of the data is divided into intervals of equal width¹ (often called bins) and the number of observations in each interval is counted. The histogram is the plot of bars, one for each bin, with heights

¹ The intervals can be of unequal width but this leads to complications and is best avoided

proportional to the number of observations in the corresponding bin. The height can be the number in each interval but the number expressed as a proportion of the total number of observation in the sample can give a picture of the distribution of the data that is not dependent on sample size.

The analyst must exercise judgment in the choice of number of bins. Figure 2 shows the data on the heights of 99 boys using different numbers of bins. With just two bins there is very little indication of how the heights are distributed. With six bins matters are clearer but a better picture is probably given by the case with 15 bins. When the number of bins increases further there is too little smoothing of the data and the histogram start to look rather jagged.

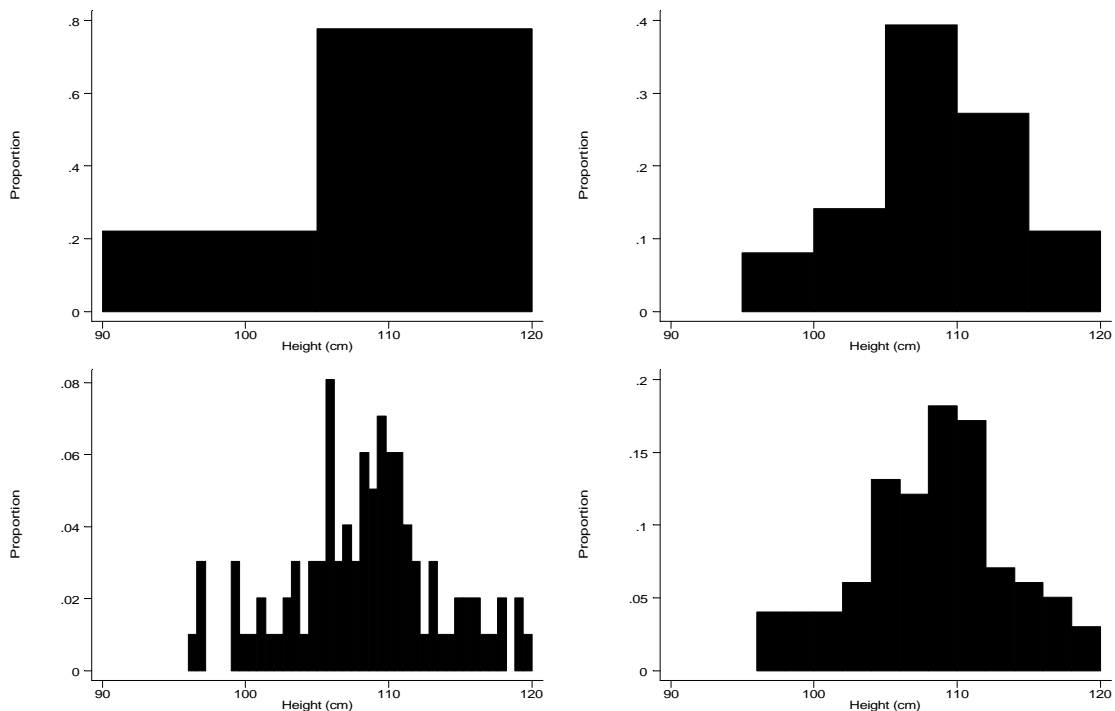


Figure 2: histograms with 2, 6, 15 and 50 bins (clockwise from top left)

Unlike box and whisker plots, which only attempt to provide a graphical version of the five number summary, histograms show the entire sample and allow the full distribution of the sample to be viewed (to within the effects of the chosen bin width). Also, as the size of the sample increases deeper aspects of the nature of the distribution may become apparent. Figure 3 shows histograms for the above sample of 99 heights together with histograms for three (simulated) larger samples, of sizes 300, 1000 and 10000.

It can be seen that as the sample size gets larger the form of the distribution becomes more regular and seems to approach an idealised form depicted by the curve, which has been superimposed on the last two histograms.

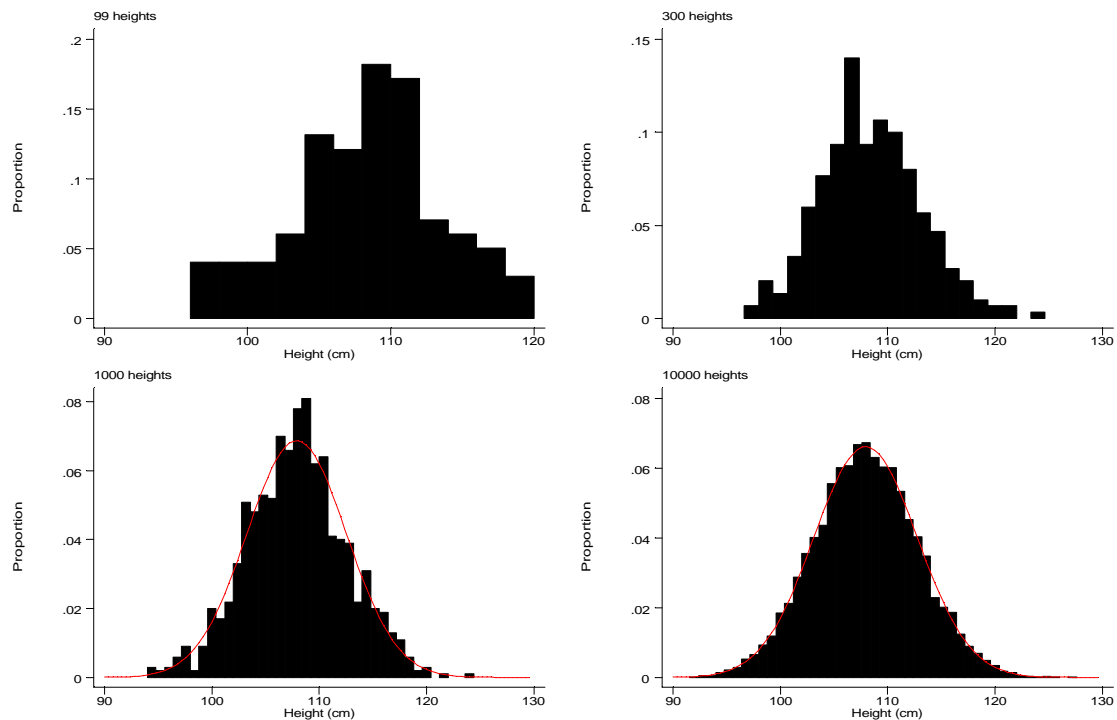


Figure 3: histograms for 99 heights and simulated samples of 300, 1000 and 10000 heights

Here the use of histograms has revealed that in very large samples a certain regularity of form, to which it appears the shape of histograms tend as the sample size increases. The form shown in the figure is symmetric and bell-shaped and in this case actually represents the well-known Normal distribution. Not all measurements will tend to distributions with this shape, although many do. Incidentally it should also be noted that the form accords closely with the hypothetical form only for a very large sample, an observation which has implications for assessing distributional form in smaller samples.

However, before the notion of a Normal distribution can be fully explained, the notion of a *population* must be introduced, and together with it the basic ideas of *inferential statistics*.

Populations, Parameters and Inference

The data on heights of 99 boys in the above example may well be of interest in their own right, but more often they are of value for what information they contain about a larger group of which they are *representative*. Similarly when conducting a clinical trial, the results obtained from the patients in the trial are, of course of value, for those involved, but they are much more important to the medical community at large for what they tell us about the group of all similar patients. In other words there is usually considerable interest in making *inferences* from the particular of the sample to the more general group. It is for this reason that much of statistics is built around a logical framework that allows us to determine which inferences can and cannot be drawn.

The important component of inferential statistics is the idea of a population. This is the group about which we wish to learn but which we will never be able to study in its

entirety. Populations are often perform conceptual: for example, the above data may come from a study in which aim is to learn about the heights of males at school entry in the UK.

A second important component of inferential statistics is the idea of a *sample*. Although we can never study a whole population, we can select several individuals from the population and study these. The selected individuals are know as a *sample*. The hope is that by studying the sample we can make inferences about the population as a whole, and it is this process which is the central concern of modern statistics. The way the selection is made must ensure that the sample is representative of the population as a whole. The main tool for ensuring representativeness is random sampling. If some process supervenes which makes the selection unrepresentative then the sample is often referred to as *biased*.

Most of the subject of statistical design is concerned with methods for selecting samples for study in a way that ensures inferences about a relevant population will be valid. A further issue is deciding how large a sample needs to be in order to attempt to ensure that the inferences will be useful.

Populations are not samples and therefore cannot be pictured by histograms. The natural way to describe how a measurement is distributed through a population is by means of a suitably defined curve, such as that superimposed on the histograms in figure 3. For continuous measurements, such as heights, weights, serum concentrations etc. by far the most important distribution is the Normal distribution and an example is pictured in figure 4.

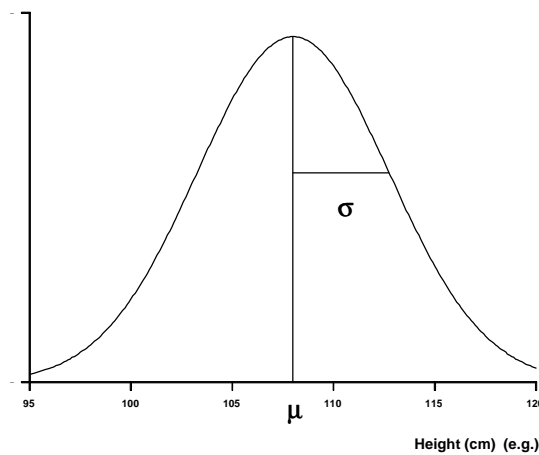


Figure 4: a Normal distribution

A Normal distribution has a central peak with the curve descending symmetrically on either side: for obvious reasons the curve is often described as being ‘bell-shaped’. The height of the curve indicates that most values in the population fall near the central value, with fewer values further from the centre. The decline is symmetric, so there will be equal amounts of the population located at the same distance above the peak as there is at that distance below the peak.

All Normal distributions have the same basic shape but they may have different locations and different spreads or *dispersions*. The location is determined by the position of the central peak and the dispersion by the width of the bell. These two attributes are determined by two *population parameters*: the peak is located at the *population mean* \boldsymbol{m} and the width is determined by \boldsymbol{s} , the *population standard deviation*. Most populations are described in terms of a distributional form together with a few population parameters. The mean and standard deviation are the only parameters that are needed to determine a Normal distribution. Other kinds of distributions may be specified in terms of other kinds of parameters.

Population parameters are unknown, as they are defined in terms of the whole population, which we can never study. There are sample analogues of population parameters and these can be estimated from the data in the sample. A fundamental idea in statistical inference is that we use these sample analogues, known as *sample statistics*, to *estimate* the corresponding population parameters. In statistical analyses it is important to distinguish clearly between population parameters, which we can never know but about which we wish to learn, from sample statistics which we can compute. To help maintain this distinction there is a widely used convention which reserves Greek letters for population parameters, such as \boldsymbol{m} and \boldsymbol{s} , and the corresponding roman letter for the corresponding sample statistic, such as m and s .

Sample Statistics for a Normal Distribution

Given a sample of data from a Normal distribution, the sample mean and sample standard deviation (sample SD) can be calculated. The sample mean² is what is loosely referred to as the ‘average’ and is found by adding up all the numbers in the sample and dividing by the number of values that have been added. As well as being mathematically the right thing to do, it is also a common-sense way to arrive at a typical value. In mathematical notation this is written as:

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

where x_1, x_2, \dots, x_n represent the individual observations in the sample. In a similar notation the sample SD can be written as:

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

For the sample of 99 heights the mean is 108.34 cm and the sample SD is 5.21 cm. Note that units should be given for *both* the mean and the SD.

In general there is little need these days to work directly with either of these formulae, as the computations will be done by computer, but however they are executed they are

² Strictly the *arithmetic mean*

fundamental to inferences for Normally distributed data. Those interested can consult appendix 2 for an explanation of why the SD is computed in the way described above.

Means and SDs are important if the data follow a Normal distribution. Ways of assessing whether data are Normal and the extent to which such assessment is necessary are discussed below. If data are found not to be Normally distributed then are means and SDs inappropriate? It is not easy to give a general answer to this question. Use of medians and quartiles will offer a correct alternative so if there is doubt means and SDs can be avoided. Certainly, if the data have a skew distribution, as illustrated in figure 5, then the mean can be alarmingly sensitive to the values of a few large observations in a way that the median is not. However, this is not a simple question and being too ready to opt for the use of medians and quartiles can lead to inefficient and unnatural analyses. There are many methods for dealing with non-Normal data, such as transforming data and using forms of means other than the arithmetic mean and these approaches can be far preferable. There will be further material on this point later in the course.

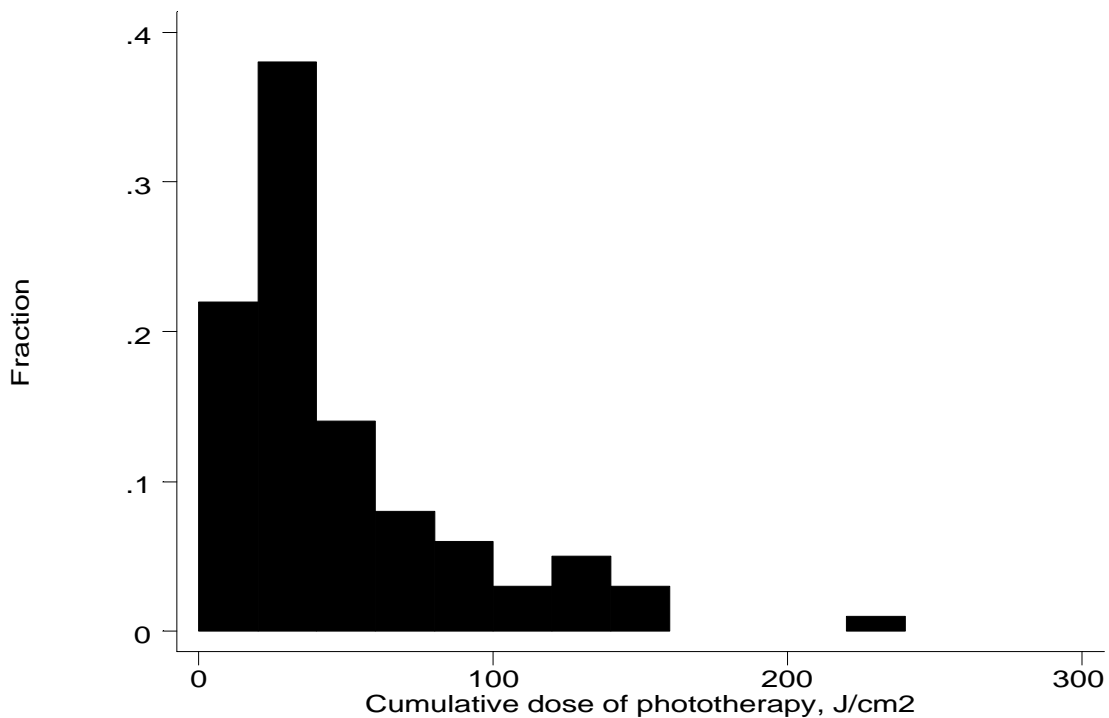


Figure 5: data having a skewed distribution (data from Dr PM Farr)

Quantitative use of the Normal curve

If the observations in the population follow a Normal distribution then this can be characterised by the population mean and SD, m and s . These can be estimated by the sample mean and SD, m and s . The shape of the Normal curve gives an impression of how the observations will be distributed but can anything more quantitative be made of the fact that the observations have a Normal distribution? Essentially the answer is

that knowledge of the values of m and s allows all aspects of the distribution to be described.

The main thing to realise about the Normal curve is that the aspect which is easiest to interpret quantitatively is the area under the curve. Figure 6 shows a Normal distribution with population mean and SD of 108 cm and 4.7 cm respectively. The area under the curve up to 112 cm has been shaded and is interpreted as the *probability* that an individual from this population has a height below 112 cm. In general the area up to a point X is P and necessarily the area under the whole curve is 1.

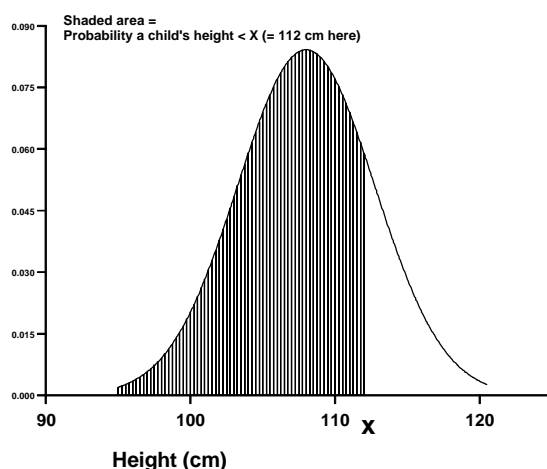


Figure 6: interpretation of area under a Normal curve

The value of P corresponding to a given X is the *cumulative probability* of the distribution at X , or the *cumulative distribution function* at X . Unfortunately there is no simple formula for this, which would allow P to be determined from X or vice versa. Computer packages will return the value of P for a given value of X if values are entered for X , m and s . If MINITAB is used and the 'Normal' sub-menu is chosen from the 'probability distributions' part of the Calc menu, the population mean 108, the population SD, 4.7, can be entered along with the target value 112 as the 'input constant'. Selection of the Cumulative probability option returns a value 0.8026, which is the value for P .

The calculations are not restricted simply to the probability of being less than a particular value. Clearly the probability that a value is above X is $1-P$. Also the symmetry of the Normal curve can be exploited: for example if Y is so many units below the mean and there is a probability Q of having a value below Y , then the probability of being above an equal number of units above the mean is also Q , as illustrated in figure 7. The probability of being between any pair of values can be found by working out the probability of being below the lower value, the probability of being above the upper value and then adding these two values and subtracting them from 1.

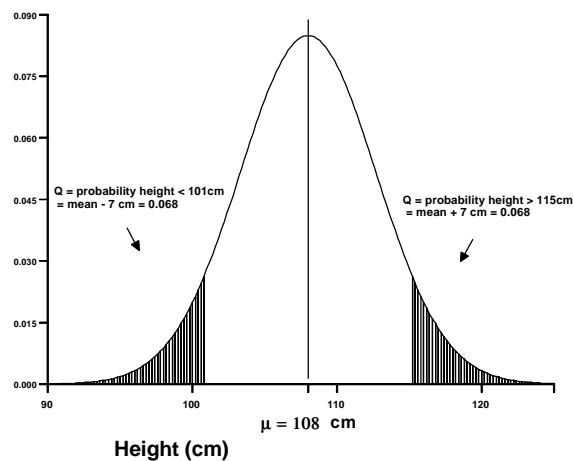


Figure 7: use of symmetry to find probabilities above as well as below given values

A further use of the Normal curve is to ask questions such as ‘what is the height such that only 3% of boys are shorter than that value?’ i.e. given a value of P what is the corresponding X ? This also needs to be computed in a statistical package. In MINITAB the same method is used as described above but selecting the *inverse cumulative probability* option. So, for example, using the mean and SD used previously it can be found that only 3% of boys have a height less than 99.16 cm (n.b. the 3% must be entered as 0.03).

For each X there is a corresponding P , but things can be made simpler than that. If the value X is written as a multiple of s above or below the mean, then P depends only on that multiple, not on the values of m and s . Writing $X = m + Zs$, this assertion states that P depends only on the value of Z . So, for example if $X = 112$ cm, as in the above example, then $112 = 108 + Z \times 4.7$, so $Z = (112 - 108) / 4.7 = 4 / 4.7 = 0.85$. If X had a Normal distribution with different mean and SD, say 100 cm and 3.5 cm, then the proportion of this population below $100 + 0.85 \times 3.5 = 102.975$ cm is the same as the proportion of the original population below 112 cm, i.e. 0.8026.

What is the advantage of knowing that P depends only on the Z value? After all, if you want to compute P for a given X you can usually tell the program values for the mean and SD. The value is actually that knowing a few pairs of Z and P values can be a considerable aid to understanding how your data are distributed. Once you know (or have estimates of) the mean and SD you can readily compute what proportion of the data is below certain values, at least for the pairs of Z and P you know. Some useful values are shown below.

| | | | | | | | |
|--|-----|------|-------|--|-------|-------|-------|
| Z | 0 | 1 | 2 | | 1.96 | 0.675 | 2.58 |
| P | 0.5 | 0.84 | 0.977 | | 0.975 | 0.75 | 0.995 |
| Proportion within Z SDs of mean ($=2P-1$) | 0 | 0.68 | 0.954 | | 0.95 | 0.5 | 0.99 |

Note that if the proportion falling below $Z (>0)$ is P , then the proportion above Z is $1-P$ and by the symmetry of the Normal curve the amount falling below $-Z$ is also $1-P$ (cf. Figure 7), so the amount below Z but above $-Z$ (the proportion within Z SDs of the mean) is $P-(1-P) = 2P-1$.

Using these pairs of Z and P it can be seen that 68% of the population is between $m-s$ and $m+s$. About 95% lie between $m \pm 2s$: in fact 95.4% fall in this interval and a more accurate interval containing 95% of the population is $m \pm 1.96s$. Intervals containing 95% of some population are widely used in statistics and if the population is, at least approximately, Normal then this corresponds to one of $m \pm 1.96s$ or $m \pm 2s$. Strictly the former should be used but the convenience of the latter means that the two versions are used almost interchangeably.

Also, the above shows, fairly obviously, that the median of the Normal distribution is m and, less obviously, the upper quartile is $m+0.675s$ (and hence the lower quartile is $m-0.675s$). Indeed, for any proportion P the value cutting off the bottom $100P\%$ of the population, often referred to as the $100P$ th *percentile* or *centile* is $m+Zs$ and this can be estimated by $m+Zs$.

From the above observations on the value of the quartiles in a Normal population the Inter-quartile range (IQR) for a Normal distribution is $(m+0.675s) - (m-0.675s) = 1.35s$. A crude and inefficient but valid way to estimate of s would be to compute the quartiles of a sample, as described in appendix 1, find the IQR and divide it by 1.35. What would be the corresponding result for the range (i.e. the maximum - minimum values)? The answer is that the range is expected to be proportional to s but the divisor analogous to 1.35 for the IQR is no longer a constant but depends on sample size. A moment's reflection will show that is sensible: if you draw a sample of size 10 from a population you would expect the extremes of that sample to be substantially less extreme than the extremes from a sample of a 1000. The conclusion is that the range of a sample not only reflects the spread of the measurements, it also reflects the sample size. As such it is seldom a useful measure or estimate of a population attribute such as spread.

Why is the Normal distribution important?

The Normal distribution is important in statistics for different reasons, although in some circumstances these reasons do overlap somewhat.

- i) it arises empirically: many variables have been studied intensively over the years and have been found to be Normally distributed. If a variable is Normally distributed then it is important to take advantage of this fact when devising and implementing statistical analyses. Moreover, many of the statistical techniques that require data to be Normally distributed are *robust*, in the sense that the properties

of these methods are not greatly affected by modest departures from Normality in the distribution of the data.

- ii) Some variables are Normally distributed because, or partly because of the biological mechanism controlling them. Variables under polygenic control are an example of this and this is illustrated in appendix 3.
- iii) The Normality arising from polygenic control is related to another reason why Normality is important. This is that the distribution of the mean of a sample tends to have a distribution close to a Normal distribution even when the individual variables making up the mean are not Normally distributed. The departure from Normality decreases as the sample size increases.

By way of illustration of this, figure 8 shows four histograms, with the best fitting Normal distribution superimposed. The first shows a sample of 10000 from a population that has a non-Normal distribution (in this case highly skewed). This illustrates the underlying distribution of the variables. The remaining histograms are each of 500 sample means: in the first each sample is of size 10, the second of 50 and the third of 100.

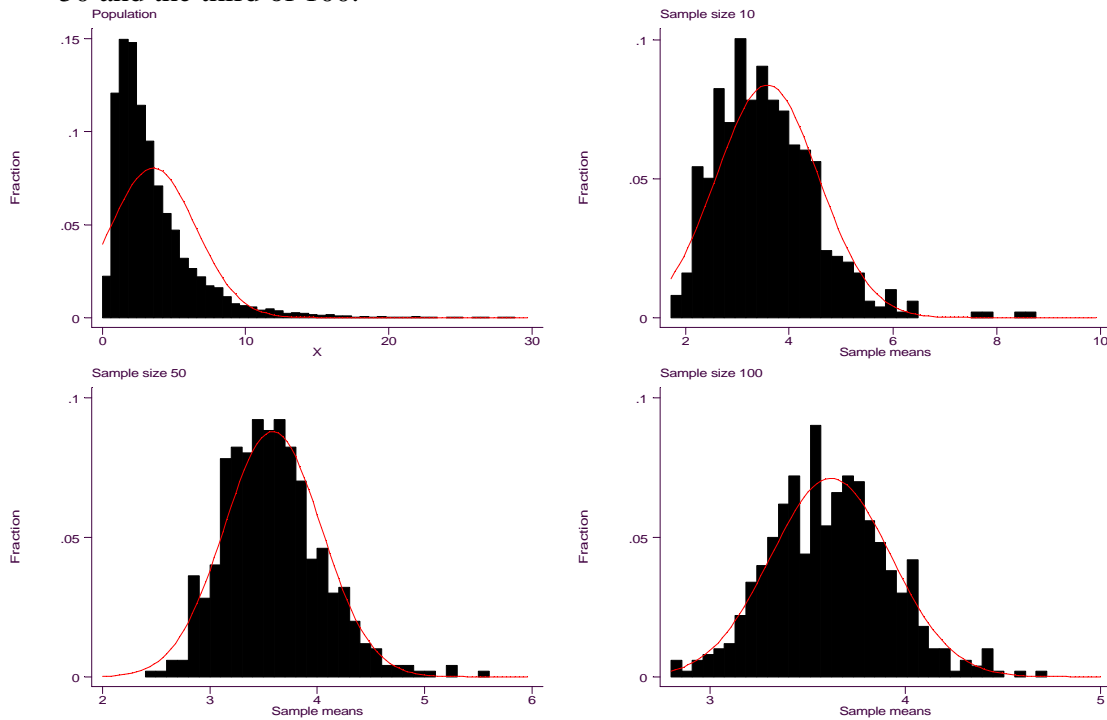


Figure 8: histograms of 10000 observations form a population and of 500 sample means for samples of size 10, 50 and 100.

It is seen that even for a population that is very skew, sample means have distributions that are much more symmetric and that are much closer to Normal than the population distribution. The effect is noticeable even in samples as small as 10.

Assessing Normality

Many statistical techniques assume that the data to which they are applied follow a Normal distribution. Many analysts are therefore anxious to determine whether or not

their data are Normally distributed so that they will know if it is legitimate to use the methods that require Normality. Some ways of doing this are described below but before going into details several general points ought to be made.

Non-parameteric or Distribution-free methods

A set of methods is available that perform a range of statistical techniques, such as performing hypothesis tests, constructing confidence intervals, that do not require an assumption of Normality. Workers therefore might wonder why it is necessary to bother with methods that make troublesome assumptions when such assumption-free approaches are available. Indeed, distribution-free methods are widely encountered in the medical literature.

While distribution-free methods can occasionally be useful, there are several reasons why a tendency to opt too readily for distribution-free methods should be resisted.

- i) In many common applications the assumption of Normality is not all that crucial, and may not even be the most important of several assumptions underlying the method. For example, the unpaired t -test for comparing two groups assumes the data in each group is Normal and has a common SD. However, quite marked departures from Normality can be tolerated and unequal SDs is probably a more serious violation of the specified conditions.
- ii) Related to this is the fact that for some important purposes, such as constructing confidence intervals, distribution-free methods are certainly not assumption-free. E.g. in the unpaired two-sample comparison the distributions are assumed to have the same shape and differ only by a shift.
- iii) Distribution-free methods are at their best for hypothesis tests and their reliance on the ranks of the data can make some aspects of estimation seem unnatural.
- iv) Distribution-free methods, at least those widely available, cannot cope with data that has anything more than the simplest structure.
- v) Transformations can be useful in changing the distribution of data to a more manageable form: an example of this is given later in the course

For many purposes it is therefore misguided to be too pre-occupied with the accurate assessment of the Normality of data. This remark is certainly reasonable when the data is to be used for things such as tests on means and confidence intervals for means. There are exceptions, when the inferences are based more strongly on the assumption of Normality. This typically arises when trying to estimate features of the data that depend on the extremes of the distribution. For example, if an estimate of the third centile of a population is required from a sample of 200 heights, the sample can be sorted and the 6th largest value chosen. However, if the sample is from a Normal distribution a much more precise estimate is $m-1.88s$: in fact the precision of this is about the same as that found by the sorting and counting method from a sample of size 800. In other words the assumption of Normality has saved you from having to measure 600 heights. However, $m-1.88s$ is only a valid estimate of the 3rd centile if the Normality assumption holds. Moreover, as the estimate is concerned with the location of points in the tail of the distribution then the validity of the estimate is much more

dependent on the distribution being very close to a Normal distribution than is the case for, say, a *t*-test.

Relative size of mean and SD: a quick and dirty test for positive variables

A feature of the Normal distribution is that it describes data that are distributed over all values, both positive and negative. If the variable so described can take negative as well as positive values then this presents no problem, either conceptual or practical. However, many variables encountered in medical science can only be positive, e.g. heights, bilirubin concentrations, haemoglobin concentrations cannot be negative. It would be too strict to deem that the Normal distribution could not be used to describe such variables, because this would preclude many useful analyses. If the Normal distribution used to describe a necessarily positive outcome ascribes very little probability (say, less than 1 to 5%) to negative values, then there will be no practical problems, even though the conceptual difficulty may remain. However, if the Normal distribution fitted to a positive variable ascribes substantial probability to negative values then this indicates that the use of Normal distribution may not be appropriate for this variable.

Above it was noted that 16% of a Normal population falls below $m-s$ and 2½% below $m-2s$. So, if a positive variable that is assumed to be Normal has an estimated mean, m , less than the estimated SD, s , i.e. $m/s < 1$ then a Normal distribution may well be inappropriate. As the ratio m/s increases towards 2 the suspicion that Normality is inappropriate will diminish. Of course, the shape of the distribution could make the use of a Normal distribution inappropriate whatever the value of m/s : this is far from a foolproof assessment, it is simply a useful rule of thumb. It can be particularly useful when reading the literature, where it may be the only method that can be applied using the published material.

Histograms and other graphical methods

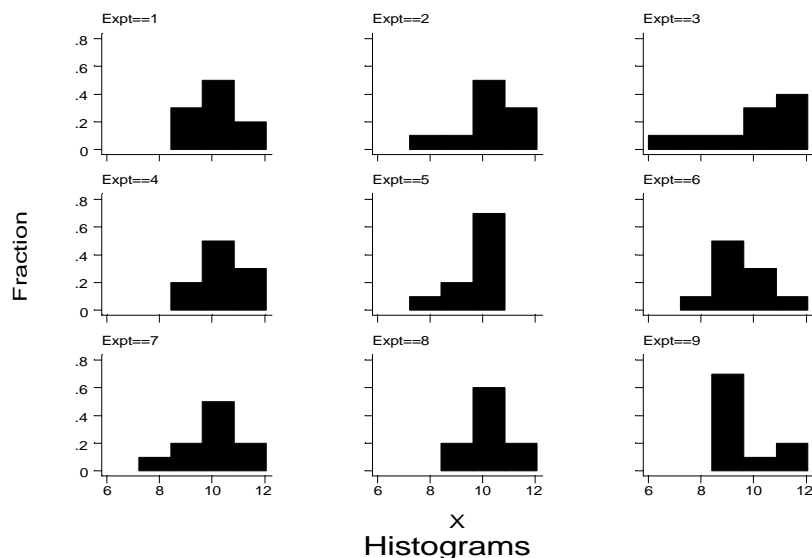


Figure 9: histograms of samples of size 10 from a Normal distribution

One of the simplest and most obvious ways to assess Normality is to plot the histogram, as was done in figure 3. This allows the general shape of the distribution to be estimated (and, incidentally, allows reasonable assessment of possible outlying values in the sample). However, some care is needed in the assessment because too strict an insistence on a suitable ‘bell-shaped’ appearance can lead to genuinely Normal data being dismissed as non-Normal. This is especially true of small samples. As an extreme example figure 9 shows histograms of nine samples, each of size 10, randomly generated from a Normal random number generator on the computer - so the data are *known* to be Normal. It is seen that several of the plots look far from Normal, emphasising that recognition of Normality, at least in small samples, is very difficult.

A slightly more sophisticated approach is to use a *Normal probability plot*. This can be obtained in Minitab (version 11) by choosing **Probability plot ...** from the **Graph** menu. Selecting the Normal distribution button and inputting the column containing the sample allows the Normal probability plot to be obtained. The default is to have confidence intervals on the plot but these are an unnecessary distraction and can be omitted by removing the tick from the appropriate box in the **Options...** box. Applying this to the data on heights of 99 boys gives:

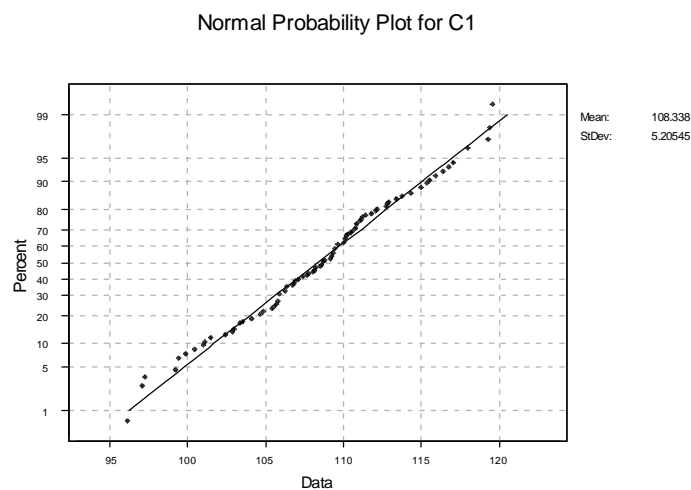


Figure 10: Normal probability plot of heights of 99 boys

The idea behind this method can be explained as follows. If a sample of size 1 were available from a Normal population then you would ‘expect’ this value to be m . If the sample were of size two you would expect the larger value to be a little above m and the smaller to be the same amount below m . If the sample were of size three the smallest of the values would be expected to be below m (by a slightly greater amount than in the sample of size two, as smaller samples would be expected to throw up less extreme maxima and minima), the largest value would be the same amount above m and the middle value would be near m . As the sample size increases, the smallest value is ‘expected’ to fall further and further below m , with the middle of the sample being close to m . Intermediate values will fall in-between, with values clustering more tightly near the mean as their rank increases, because of the nature of the bell shape of the Normal curve. The largest values will be arranged above m symmetrically relative to the smallest values. This idea is illustrated in figure 11 for increasing sample sizes.

If a sample is from a Normal distribution then if it is sorted into ascending order and each point is plotted against the 'expected' position of that point in the ordering, a straight line should result. Of course, sampling variation means that the line will never be exactly straight and some judgment must be used in the interpretation of these plots. That in figure 10 is more than satisfactory: an example of a Normal plot of a non-Normal sample is shown in figure 12 and is clearly far from straight: the data are a sample of 99 from the population used in the first panel of figure 8.

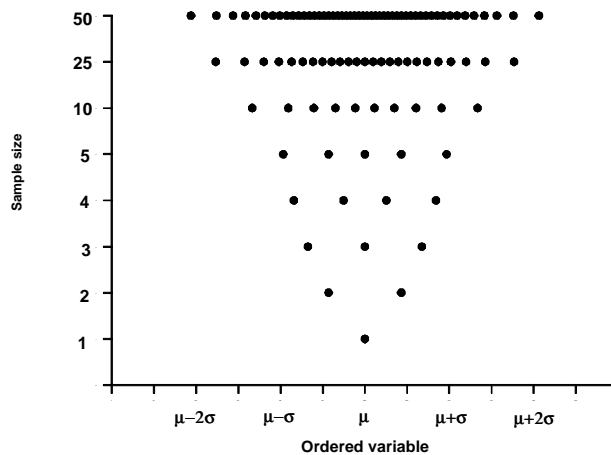


Figure 11: expected locations of ordered elements of a sample from a Normal distribution for various sample sizes

The labelling of the graphs in figures 10 and 12 needs a brief word of explanation. The ordered sample values are plotted on the horizontal axis and the vertical axis is the corresponding expected value. However it is not labelled as such. The scale is labelled as a percentage and this is best explained by an example. If the sample has 100 points then the smallest is plotted at about 1%, the next at about 2% etc.

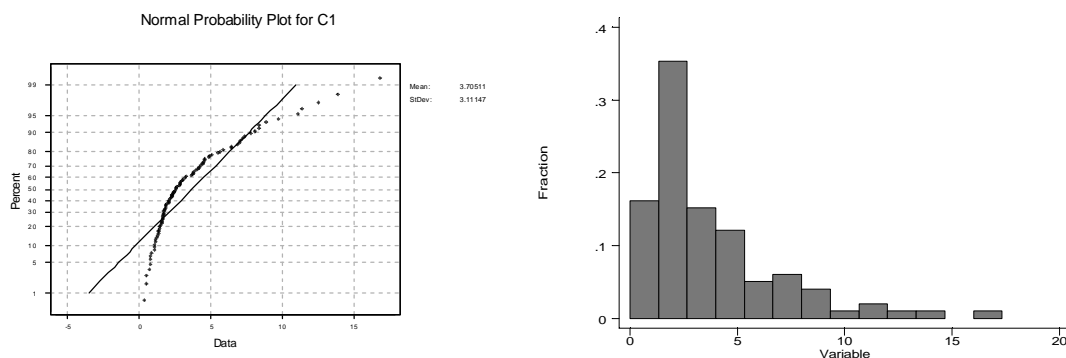


Figure 12: Normal probability plot and histogram for a grossly non-Normal sample of size 99

There are other methods for assessing Normality, including formal hypothesis tests but these generally lack power. The Normal probability plot is one of the best and, in experienced hands, one of the most useful ways to assess Normality.

Appendix 1: computing medians and quartiles

Once the sample has been sorted into ascending order the median is the middle value. More precisely, if the sample contains n values then the median is the $\frac{1}{2}(n+1)$ th largest value. In the example $n=99$, so $n+1=100$ and the median is the 50th largest value. Had the sample been of size 100, for example another child, of height 118.1cm, been measured, then the median would have been the $\frac{1}{2}(100+1)$ th = 50½th largest value. Of course, there is no 50½th largest value until we interpret what is meant by a fractional rank. In this augmented sample the 50th largest value is 108.7 cm and the 51st largest value 108.8 cm: the 50½th largest value is interpreted as being ½ way between these values, i.e. the median of the augmented sample is $108.7 + \frac{1}{2} (108.8 - 108.7) = 108.75$ cm.

The definitions for the quartiles follow by analogy. The lower quartile for a sample of size n is the $\frac{1}{4} (n+1)$ th value and the upper quartile is the $\frac{3}{4} (n+1)$ th largest value. In the example above where $n+1 = 100$, the lower quartile is the 25th largest value and the upper quartile is the 75th largest value. This definition could result in fractional ranks of $\frac{1}{4}$ and $\frac{3}{4}$ which are interpreted in the same way as above. In the augmented sample of size 100, $n+1 = 101$ and lower quartile is the 25¼ th largest value: this is ¼ of the way from the 25th to the 26th largest value. The 25 th largest value is 105.6 cm and the 26 th largest value is 105.7 cm, so the lower quartile is $105.6 + \frac{1}{4} (105.7 - 105.6) = 105.625$ cm.

Appendix 2: the form of the SD

The idea in computing a measure of spread is first to define a measure of location and then to measure how dispersed the observations are about that point. In this instance the mean is the measure of location, so it is natural that the SD should be based on the differences of the individual points from the sample mean, i.e. the expression is based on $x_1 - m, x_2 - m, \dots, x_n - m$.

However, one ‘obvious’ measure of spread, namely to find the average of these deviations, will not work. This is because if you add up $x_1 - m, x_2 - m, \dots, x_n - m$ the result will always be zero because of the definition of the mean: there will always be the same total deviation on the negative side (below m) as there is on the positive side (above m). A simple way round this difficulty would be to find the average of these deviations without regard to their sign. This is known as the mean absolute deviation about the mean and is a legitimate measure of spread. However its mathematical properties are poor and the theory based on this measure of spread is not as rich as that using the SD, so it is seldom used.

An alternative way to remove the balance between positive and negative deviations is to square all the deviations, making them all positive and then take the ‘mean’ of these values. This results in a quantity which is known as the *variance*. It is a very important quantity in theoretical work but is of little direct use in practice because it is on the scale of squared units, i.e. the variance of the heights would be in cm^2 . This is readily rectified by taking the square root, and it is this quantity which is the SD, and which has the same units as the original measurements.

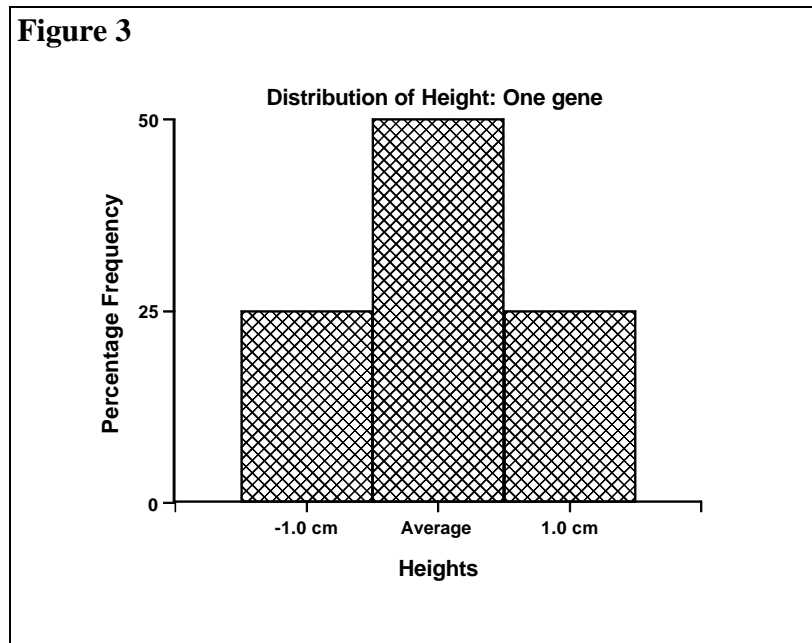
The ‘mean’ of the squares deviations was placed in quotes because, instead of dividing the sum of the n squared deviations by n , the divisor $n-1$ is used. This is largely a technical matter and is related to the fact that the spread would ideally be measured about the population mean μ rather than the sample mean m . The sample mean, being determined as, in some sense, the ‘centre’ of the sample is actually slightly more central to the sample than the population mean. Consequently the dispersion around the sample mean is slightly less than it would be about the population mean. A correction which slightly inflates the SD is therefore appropriate and is achieved by dividing by $n-1$ rather than n .

This description has considered various aspects of the formula for the sample SD. It has not covered why the quantity defined as the sample SD is the appropriate estimator of the parameter which governs the width of the Normal curve. It is, but a demonstration of this is beyond the scope of the present discussion.

Appendix 3: an illustration of how a Normal distribution may arise.

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

| | | Gene 1 | | | |
|--------|-----------|-----------|-----------|-----------|-----------|
| | | <i>hh</i> | <i>hH</i> | <i>Hh</i> | <i>HH</i> |
| Gene 2 | <i>hh</i> | -2cm | -1cm | -1cm | 0cm |
| | <i>hH</i> | -1cm | 0cm | 0cm | 1cm |
| | <i>Hh</i> | -1cm | 0cm | 0cm | 1cm |
| | <i>HH</i> | 0cm | 1cm | 1cm | 2cm |

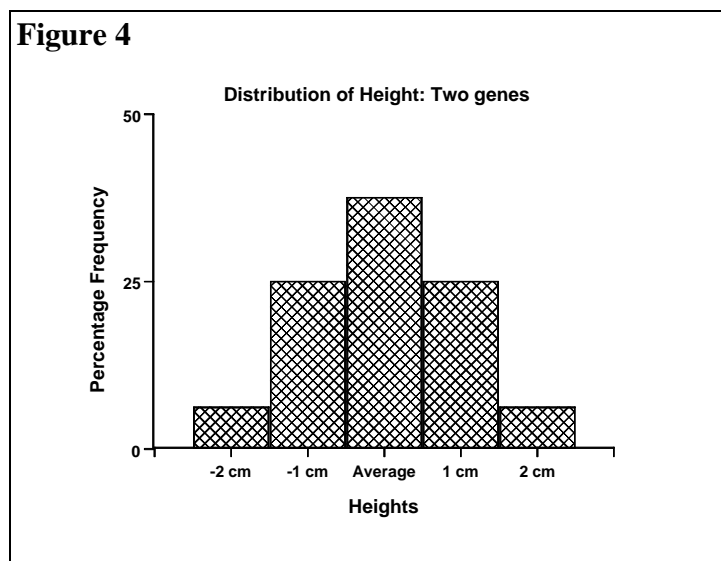
homozygote (a clinically important example is sickle-cell anaemia[§]).

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

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

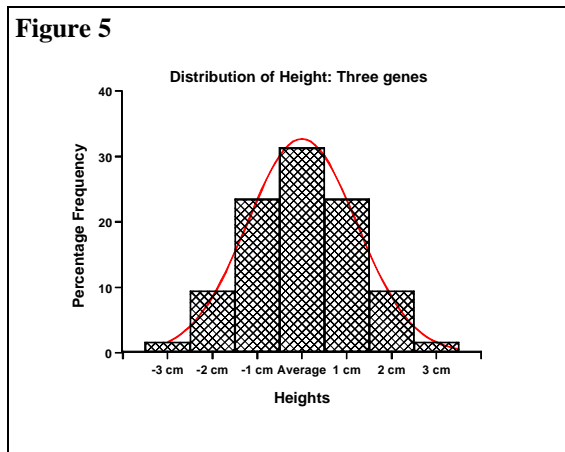
below average (hh) and 1 cm above average (HH). If the alleles H and h are equally prevalent each combination HH , hh , Hh and hH is also equally likely (where hH and Hh have been used to distinguish the heterozygote where h comes from, respectively the mother or father). However, Hh and hH both have 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 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.

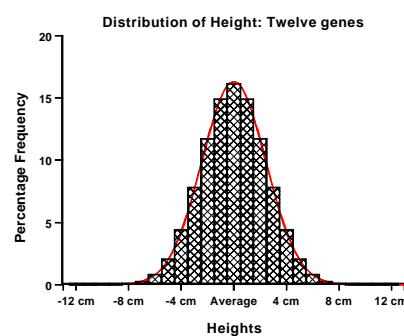
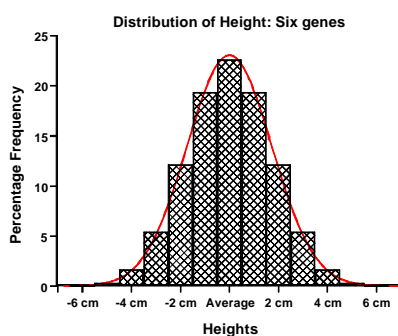
If the number of genes controlling height is now supposed to be 3, there are $4 \times 4 \times 4 = 64$ equally likely gene combinations, but these give rise to only seven phenotypes, namely heights at 1cm intervals from -3cm to 3 cm.



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.



Figures 6a and 6b

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.