Research Methods 2

Week 4: Document 1

Populations and Parameters

Populations

Last week you learnt how to summarise a set of data. The examples you encountered were the heights of 99 five-year-old boys and the haemoglobin concentrations recorded on 35 healthy males. Why should you want to do this? The obvious answer, given last week, is that the whole sample constitutes a rather indigestible mass and a summary in terms of, say, five or fewer numbers or using a graph, makes matters more manageable. However, this answer is, in some ways, rather superficial. Why are the heights of these 99 boys, or the haemoglobin concentrations of the 35 males of particular interest? While there may be circumstances where a particular set of data is of special interest, it is more frequently the case that the data are of value for what information they contain about a larger group of which they are *representative*.

Medical research is replete with such instances. For example, 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 data to a 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 conceptual: for example, the data on the heights of boys may come from a study in which aim is to learn about the heights of males at school entry in the UK. It would not be a practical proposition to study all boys in the UK at school entry, but this does not mean that this is not the group relevant to our purposes.

The aim of inferential statistics is to attempt to prescribe what can be learnt about a population as a whole by studying small, suitably chosen parts of it. The key idea here is that 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 said to be *biassed*.

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. Later in this course we will consider the issue of sample size. However, general issues of design are beyond the scope of this course. Those interested in this very important area can find a very good discussion in Bland, chapters 1-3.

If we are to select samples so that we can attempt to describe a population, it is important to have a framework for describing populations. In week 3 we learnt how to describe a sample using a histogram. A figure you met last week is reproduced below as figure 1. It is reasonable to suppose that as the size of a sample get larger, the picture it paints of the population becomes more accurate.

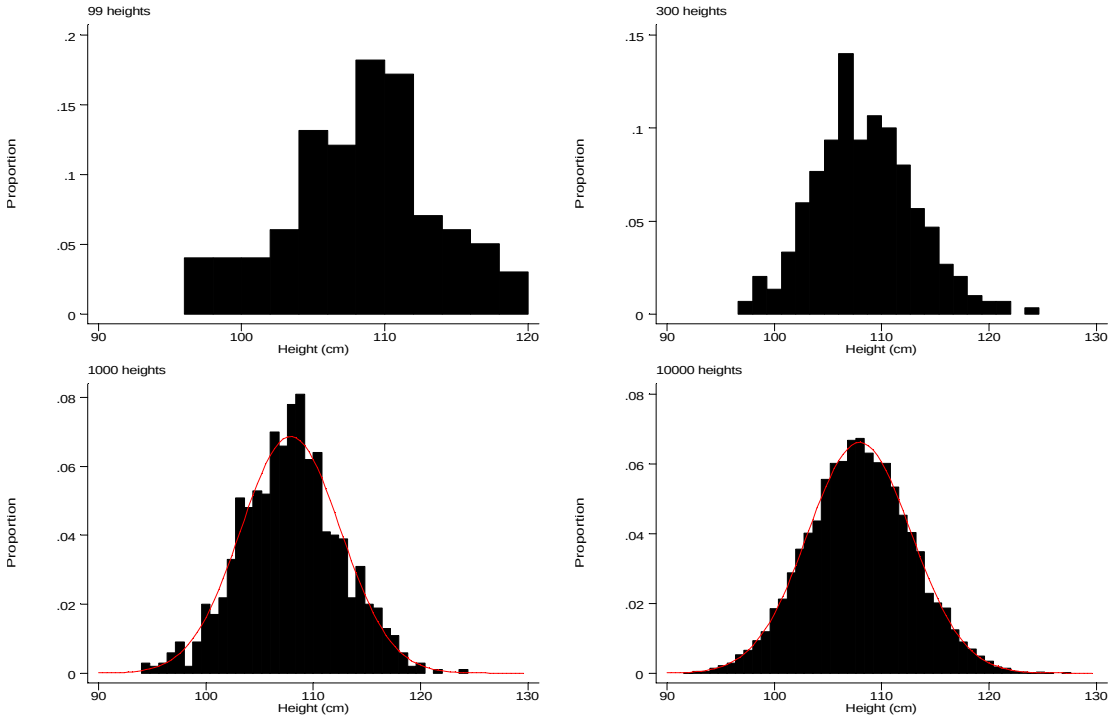

Figure 1: histograms for samples of 99 (top left), 300 (top right), 1000 (bottom left) and 10000 (bottom right) heights

It is at least plausible that as the sample size gets bigger and bigger the histograms accord more and more closely with the red curve shown in the lower panels of figure 1, and that this curve would be a sensible way of describing how the distribution of heights of five-year-old boys in the UK.

The red curve shown above is known as the Normal distribution. For continuous measurements, such as heights, weights, serum concentrations etc. it is by far the most important distribution and is the only one we shall study in detail. Outcomes which are not continuous, such as tumour stage, or whether or not a patient is in remission, which can only take one of a discrete number of values, need to be handled differently and will not be considered until week 9.

A graph depicting a population with a Normal distribution is shown in figure 2 below. The 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.

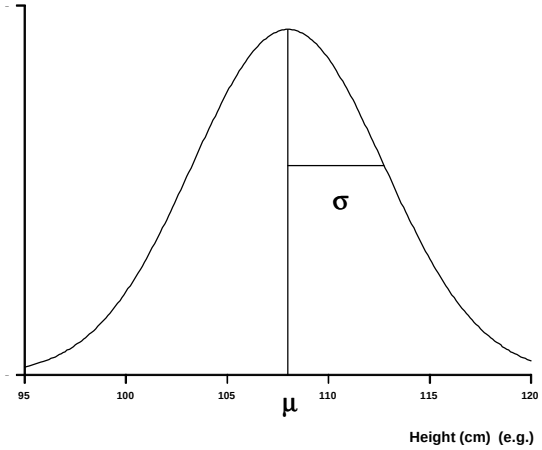

Figure 2: a Normal distribution

All Normal distribution 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* m and the width is determined by s, the *population standard deviation*.

Figure 3 below shows three Normal distributions. Each has the same population mean (here 108 cm) but different population standard deviations. The blue distribution, which has the highest peak, is the least dispersed: most of the members of this population have heights between 104 and 112 cm. It has the smallest standard deviation, namely 2cm. The red curve represents a population that is much more dispersed and has the largest standard deviation of the three populations shown of 7 cm. The black population is intermediate and has an SD (the common abbreviation for 'standard deviation') of 5 cm.

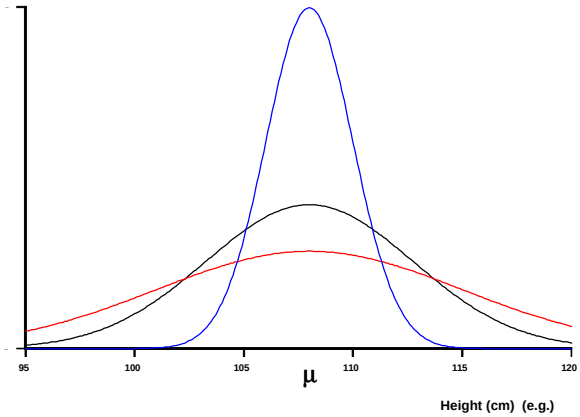

Figure 3: three populations with a Normal distribution. Each has the same mean but different standard deviations: 2cm for blue, 5 cm for black and 7 cm for red.

Most populations are described in terms of a distributional form (which here is the Normal distribution) together with a few population parameters. The mean and standard deviation are the only parameters that are needed to determine a Normal distribution. So, if we believe a variable, such as the height of five-year-old British boys, has a Normal distribution then all we need to know in order to know all about this distribution are the values of the population parameters, i.e. the mean and standard deviation.

Unfortunately, population parameters are generally unknown., so at this point the construction of the notion of an unknown population may seem a rather unhelpful exercise. However, we *can* use data to learn *something* about a population and the first steps in this direction are taken in the next document.