Research Methods 2

Week 4: Document 2

Samples and Statistics

At the end of Document 1 we were able to define a Normal population as one with a particular form (a bell-shaped curve) which was defined in terms of two parameters, the population mean, μ, and standard deviation, σ. These are almost always unknown because they are defined in terms of the whole population, which we almost never study in its entirety. Consequently this may seem to be a rather useless formalism because without knowing the values of the parameters little has been gained. However, there are sample analogues of population parameters and these *can* be evaluated from data in a sample drawn from the population. A fundamental idea in statistical inference is that we use these sample analogues, known as *sample statistics*, to *estimate* the corresponding population parameters. Of course, while we are better informed once we know the values of the statistics, they will not be exactly equal to the population parameters. The likely degree of disagreement between the statistics and the parameters is a major preoccupation of statistical science, which we will start to consider in week 6.

> 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 m and s, and the corresponding roman letter for the corresponding sample statistic, such as *m* and *s*.

Sample Statistics

The Sample Mean

Given a sample of data from a Normal distribution, the sample mean and sample standard deviation (sample SD) can be calculated. The sample mean[1] 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:

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

[1] Strictly the *arithmetic mean*

where $x_1, x_2, \ldots, x_n$ represent the individual observations in the sample. So, for example, suppose we had measured the heights of the first five boys in the table in Week 3: Document 1, we would have a sample *of size 5* comprising

117.9 106.3 103.3 105.9 116.7

In the above notation we have $x_1 = 117.9$, $x_2 = 106.3$, $x_3 = 103.3$, $x_4 = 105.9$ and $x_5 = 116.7$ and the sample mean, *m*,would be computed as:

$$m = \frac{117.9 + 106.3 + 103.3 + 105.9 + 116.7}{5} = \frac{550.1}{5} = 110.02 \text{ cm.}$$

This value is our *estimate* of the population mean, *m*, based on these five heights. So we are no longer completely in the dark about the value of the population mean. We do not *know* its value, but at least we have an estimate of it. The estimate may not be very good, as it is based on just five heights and commonsense suggests that the sample mean of all the 99 heights shown in Week 3: Document 1, namely 108.34 cm, might be better. Nevertheless, if the five heights were all we had, then at least we have got somewhere. It turns out that it is indeed better to use the larger sample, and the sense in which this is true is explained in Week 6.

The Sample Standard Deviation

While the notation may be new, the idea of getting the average by adding up all the numbers in the sample and dividing by the number of numbers will be fairly familiar. The idea behind the sample SD is the same – we compute a number, *s*, based on the data in the sample and use this as an estimate of σ. However, the formula needs to be rather different as it is measuringspread, not location. In a similar notation the sample SD can be written as:

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

The formula may be rather unfamiliar and possibly intimidating but, after the following brief illustration, we will never have to work again with this formula.

Using the above sample of size 5, the formula gives:

$$s = \sqrt{\frac{(117.9 - 110.02)^2 + (106.3 - 110.02)^2 + (103.3 - 110.02)^2 + (105.9 - 110.02)^2 + (116.7 - 110.02)^2}{5 - 1}}$$

Evaluating the subtractions in this formula and squaring them gives

$$s = \sqrt{\frac{62.094 + 13.8384 + 45.1584 + 16.9744 + 44.6224}{4}} = \sqrt{\frac{182.688}{4}} = \sqrt{45.672} = 6.76 \text{ cm}$$

This value is our estimate of σ. As with the mean, the sample SD based on all 99 heights, namely 5.21 cm, is actually a better estimate than 6.76 cm, as it is based on more data. Note that the standard deviation, as well as the mean, has the dimensions of height.

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 evaluated they are fundamental to inferences for Normally distributed data. Those interested can consult Appendix 1 for an explanation of why the SD is computed in the way described above.

Are all data Normally distributed?

If your data follow a Normal distribution then it follows i) that the relevant population parameters are the mean and SD and ii) that you should use the sample mean and SD to estimate these quantities. However, might there be circumstances where this is inappropriate? Briefly the answer is 'yes'. As mentioned above the Normal distribution is a distribution which describes a continuous variable. This is a variable, such as blood pressure, height, haemoglobin concentration, which can in principle take any value (possibly within a given range). Variables such as tumour stage or whether or not a patient is in remission, which can take only a few discrete values cannot be Normally distributed.

However, not all continuous variables follow a Normal distribution. For example, a Normal distribution is *symmetric*, that is the population spreads out above the central location in the same way that it spreads out below this value. In figure 1 below is a histogram of data which clearly does not have this attribute and so cannot be Normally distributed. Data with this kind of asymmetric distribution are said to be *skewed*. There are, in fact, many ways in which a continuous variable can have a non-Normal distribution but we will not consider this point further.

If a population cannot be assumed to follow a Normal distribution then you may have to be wary about using the mean and SD as summaries of this quantity. Means and SDs might be useful in this circumstance but more often you will have to make use of medians and quartiles as an alternative.

Nevertheless, the many variables *do* follow a Normal distribution. In addition there are many circumstances in statistics where the Normal distribution has an important indirect role and which make it an important distribution to understand. Next week some of the properties of the Normal distribution will be discussed in more detail.

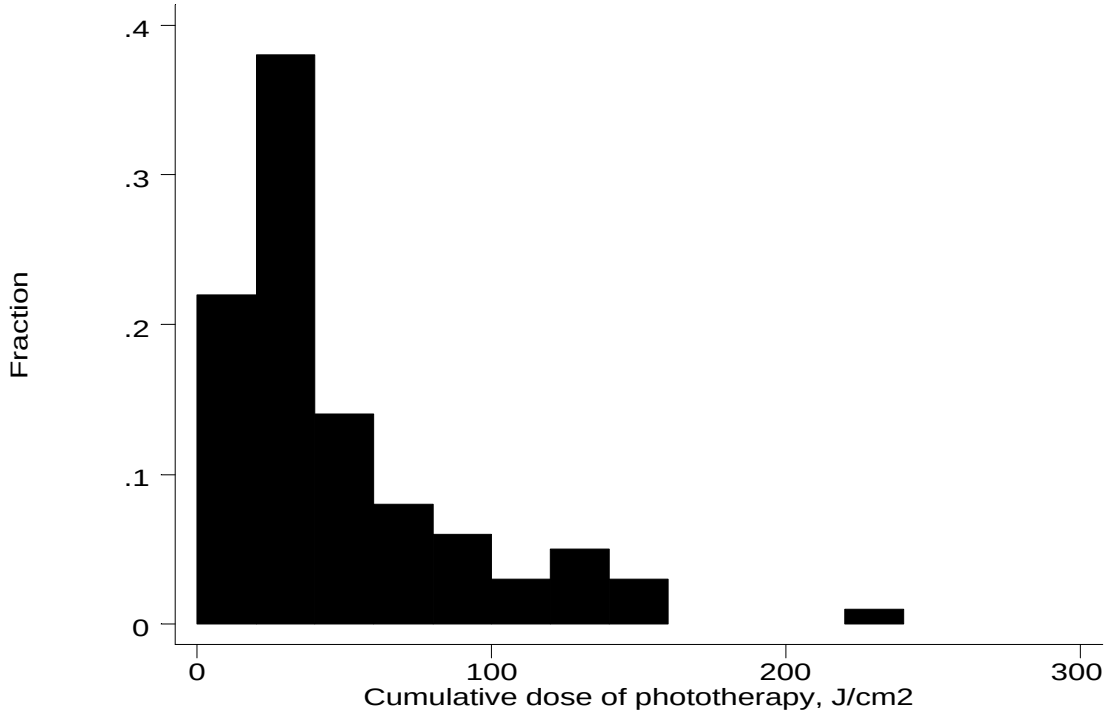

Figure 1: data having a skewed distribution (data from Dr PM Farr)