Research Methods 2

Week 5: Document 1

Quantitative use of the Normal curve

Last week we learnt that observations from many populations follow a Normal distribution. The Normal distribution was shown as a bell-shaped curve, with the location and width of the peak determined by two parameters, the population mean and SD. We also learnt that these parameters can be estimated by the sample mean and SD, *m* and *s*.

Can we be more specific about what it means for a variable to have a Normal distribution? In particular, what are the implications of the bell-shaped curve? So far we have remarked that the shape indicates that a higher proportion of the population lies under the central peak, with decreasing proportions in the tails. Can we be more quantitative than this?

The answer is yes. If we assume that the parameters m and s are known, then it turns out that we know everything about the distribution of the variable. Of course, in practice, these parameters won't be known and we would have to make do with the values from estimates *m* and *s*.

The main thing to realise about the Normal curve is that the aspect which is easiest to interpret quantitatively is not the height of the curve but the area under the curve. The area under the whole curve is one. The area under the curve up to a given value is the proportion of the population which has values smaller than the given value. This is exemplified in Figure 1, which shows a Normal distribution with population mean and SD of 108 cm and 4.7 cm respectively. If you want to have a concrete example in mind, this would be very similar to the population of the heights of five-year-old boys in the UK. The area under the curve up to 112 cm has been shaded and is interpreted as the *proportion* of the population that have heights up to 112 cm. This area is 0.8026 – in other words about 80% of five-year-old boys have heights below 112 cm.

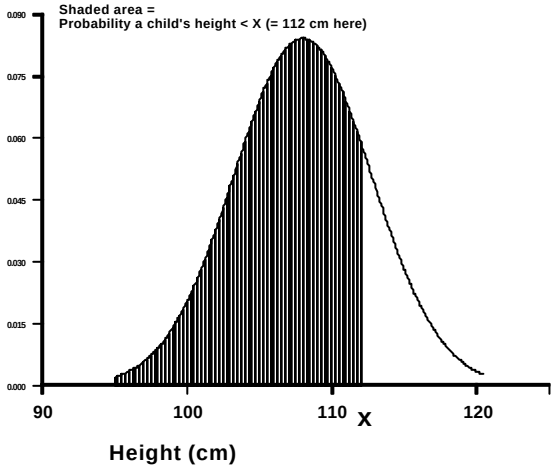

Figure 1: interpretation of area under a Normal curve

The value of the proportion, *P*, of boys with heights less than height *X* cm is known as the *cumulative proportion* (or sometimes the *cumulative probability*) of the distribution at *X* cm. Unfortunately there is no simple formula for obtaining *P* from *X* (e.g. there is not way you can obtain the value 0.8026 given above from the value of 112 cm to which it refers without the use of either a computer or tables).

A full discussion of cumulative probability is not necessary for our present purposes. However, there are three features of the Normal which it is useful to know.

1. **Use the symmetry of the distribution to extend your knowledge**.
   It may seem rather restrictive only to be interested in the proportion of boys with heights less than 112 cm. What proportion have heights above 112 cm or below 104 cm or between these heights? In fact, we already know the answers to all these questions

   If 80% of these boys have heights below 112 cm then 20% have heights above 112 cm.

   112 cm is 4 cm above the mean of the population, which is 108 cm. By the symmetry of the Normal curve about its mean, the proportion with heights above 108 – 4 =104 cm will be the same as the proportion with heights below 112 cm, i.e. 80%. Consequently the proportion with heights below 104 cm is 20%.

   If the proportion with height below 104 cm is 20%, and the proportion with height below 112 cm is 80% then the proportion with height between these values is 80 – 20 = 60%.

   The point of these calculations is to illustrate that simply by being told that the proportion of boys with heights below 112 cm is 80% a variety of different quantities can be obtained, and you are not restricted to answering just one type of question.

2. **The proportion *P* depends only on the number of SDs that *X* is from the mean**
   80% of boys have heights below 112 cm, in a population with mean 108 cm and SD 4.7 cm. The value 112 cm is nearly a whole SD larger than the mean (it is, in fact, 0.85 SDs above the mean, as $112 = 108 + 4.7 \times 0.85$ cm). So another way of putting this fact about the Normal distribution is to say that 80% of the population have height which is less than a value which is 0.85 SDs above the mean. This may seem rather a long way round to describe things but it is, in fact, very useful. The assertion that 80% of heights lie below 112 cm is not universally true, it follows because the mean is 108 cm and the SD is 4.7cm. On the other hand the second form of the statement is true for *any* Normal distribution, i.e. it is always the case that 80% of a Normal population lies below the mean plus 0.85 SD (in symbols $\mu + 0.85\sigma$).

   The fact that the proportion of a population falling below a value depends only on the number of SDs the value is away from the mean allows us to assert the 68-95-99.5 rule for the Normal distribution

3. **The 68-95-99.7 Rule**
   For a population with a Normal distribution, 68% of the population lie within one SD of the mean, 95% lie within two SDs of the mean and 99.7% of the population lies within 3 SDs of the mean. This is illustrated in figure 2 below.
   So, e.g., for the population of heights of boys with mean 108 cm and SD 4.7 cm, 68% of boys have heights between 108-4.7 = 103.3 cm and 108 + 4.7 ==112.7 cm, 95% of boys have heights between $108 - 2 \times 4.7 = 98.6$ cm and $108 + 2 \times 4.7 = 117.4$ cm, and 99.7% have heights between $108 - 3 \times 4.7 = 93.9$ cm and $108 + 3 \times 4.7 = 122.1$ cm. For a different Normal population, say one representing haemoglobin concentrations, with mean 15 g/dl and SD 1 g/dl, the same calculations can be performed, so, e.g., 95% of people in this population will have haemoglobin concentrations between $15 - 2 \times 1 = 13$ g/dl and $15 + 2 \times 1 = 17$ g/dl.
   Also, the rule can be subjected to the kinds of argument mentioned in item 1. For example, if 68% of a Normal population lies within one SD of the mean, then 32% lies outside this range and, by symmetry, 16% of the population have values that are more than one SD less than the mean (i.e. less than $\mu - \sigma$).
   The rule, especially the '95' part of it, will be used extensively in later weeks.

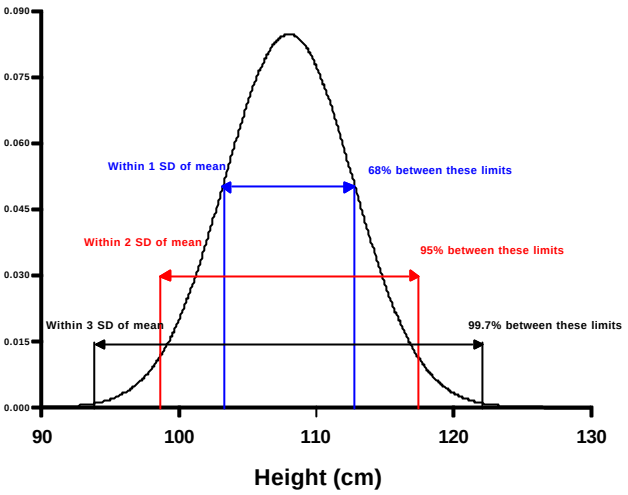

Figure 2: a Normal distribution (mean 108 cm, SD 4.7 cm) and the 68-95-99.7 rule

---

How to obtain *P* from *X* (for those who want to know, others can omit this bit)

As mentioned above, there is no simple formula which allows *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*.

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, with 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). Note that this is consistent with the 68-95-99.7 rule, which says that 2½% of the population have a height below 98.6 cm/
