Research Methods 2

Week 4: Note on generating data

It has been explained that populations, and in particular their parameters, are unknown and that we must do our best to *estimate* them on the basis of samples drawn from the population.

When learning statistical methods it often adds realism to the exercise if you analyse genuine datasets. Unfortunately, there is a snag with this. A sample of real data will be drawn from a real population, so the parameters will be unknown. As such, it will be unclear to the student whether the statistics calculated from the sample bear any relationship to the 'correct' values, namely the values of the population parameters. In addition, it is likely that real data will be obtained from a sample of a given size. Consequently, there is only a limited amount of exploration the student can do concerning the effect of sample size on the properties of sample statistics.

A way out of this difficulty is to use a facility in Minitab (which is also available in many other packages) which allows the user to specify a population with a given type of distribution (e.g. Normal) and with *known* parameter values. The user can then draw a sample of any size from this population.

So, for example, you could ask Minitab to produce a sample of size 100 from a Normal population with population mean m, = 50 and population SD σ = 5. This would be stored in a column specified by the user and could be analysed. For example, a histogram could be drawn to see how the shape of the sample compared with the bell-shaped form which is *known* to apply to the population. You could also calculate the mean and SD of the sample to see how far they are from the *known* population values.

Having done this, you could repeat the command that gave the sample and this will provide a new sample of size 100. The mean and SD of the new sample can then be computed. By doing this a few times you will soon see how sample means and SDs relate to the underlying population values they purport to estimate.

You can repeat the exercise changing the values of population parameters and also altering sample sizes, although this latter change will be considered in more detail next week.

The use of artificial data may seem rather contrived. However, the ability to 'know the answers' and to manipulate features such as sample size, which cannot be done with real data, is an important compensation. Also, we will analyse real data later in the course!

Return to Exercise Sheet 1