STATISTICS FROM THE INSIDE

1. Populations and samples

M J R Healy

This is the first of a series of notes on medical statistics. I hope to explain some of the basic ideas in a way which will be useful not only to those preparing papers for the journal but especially to those who read the papers that are published. I shall not attempt to go into details about the practical issues of analysing data as these are amply covered elsewhere; instead I shall try to convey the essence of the concepts which (like it or not) you are applying whenever you carry out a statistical analysis or interpret one in a published paper. Comments from readers would be welcome. Unfortunately, statistics is an odd subject in that the most difficult part comes first—the fundamental ideas are quite tricky to grasp, while the mechanics of bringing these ideas to bear are relatively simple. Accordingly, the first few of these notes are liable to be harder to come to terms with than the later ones.

The main objective of a statistical analysis consists in an attempt at *generalisation*. When you read a scientific paper which describes the effect of a new treatment on 20 patients, you are not usually interested in these 20 patients for their own sake; they are not your patients and they are in any case a thing of the past. Rather you want to know to what extent the reported outcomes can be generalised so that you can apply them to your own patients, both present and future. The problem of generalisation—of arguing from the particular to the general—is a very fundamental one and I start by considering the way in which statistics copes with it.

The underlying methodology of statistics—as of all science—consists in setting up a *model* of the situation at hand. Specifically, when confronted with a body of data relating to a number of patients, statistics considers a large collection of all the actual and potential patients to which you and others might wish to apply the results being analysed. This collection is known technically as a *population* (the word is a hangover from the origins of statistics in the 17th and 18th centuries). The objective of the statistical analysis will be to say something about this population. In order to do this, it is supposed that the cases reported constitute a *sample* from the population. To make generalisation possible, they must be a *representative* sample; in particular, they are almost always treated as if they were (or at least behaved like) a *random sample*.

Each member of the population will possess an *attribute* which is the subject of the study. This may be purely descriptive, such as ABO blood group, or yes/no, such as infected/not infected, or quantitative, such as family size or stature. A quantitative attribute may be a *discrete* count or a *continuous* measurement. This attribute is known as the *variate* of the population (a population may have several variates, giving rise to *multivariate* problems). A statistical analysis is aimed at saying something about the different values of the variate which occur in the population (or populations) arising in the study. The phraseology for doing this is that of *probability*. Roughly speaking, the analysis aims at telling you how probable are the different values of the variate which occur in the population, that is, at describing the way in which the total probability of 1·0 is *distributed* over the possible variate values. Often (but not always) this is done by introducing a mathematical relationship between probability and value, using some form of *theoretical distribution*. The most familiar of these in practice, at least for continuous variates, is the so called *Normal* distribution (I write the name with a capital to stress that it is a technical term; its opposite is non-Normal, *not* abnormal). Discrete data in the form of counts can often be described by the *binomial* or *Poisson* distributions. Both of these closely resemble the Normal distribution when the counts are not too small.

The statistical model, then, assumes that the cases in a particular study are a random sample from a population. When you read an account of the study, you may also suppose that your own cases are another such sample from the *same* population. Knowledge about the population (based upon the sample in the paper) will then tell you something about how often different variate values might be expected to occur in your own experience—how many cases treated with a new drug might be expected to improve, for example. It is clearly a matter of judgment whether it is reasonable to consider the cases reported in the paper and your own cases as samples from the same population. To help with this, it is important for authors to report in detail the criteria and methods by which their cases were selected.

Statistics thus has the problem of finding ways of describing a population's variate values in a comprehensible way. First, we would wish to know whereabouts most of the values lie, their *location*. Various summary measures are possible, of which the most important are the *mean*, which is the straightforward average of the variate values, and the *median*, which is the middle value when all the values are arranged in ascending order. Almost all populations which occur in practice have values which cluster around a central value, with the probability of finding more remote values decreasing fairly rapidly. The central, most 'popular' value is called the *mode*.

Once we know the central location of the

17 Moreton Avenue, Harpenden, Herts AL5 2EU

Correspondence to: Professor Healy.

variate values, the next question is their *variability*, how far away from the central value they commonly occur. To measure this, we might take each value, subtract the population mean to see how far the value deviates from it, and average all the *deviations* so formed. This does not work; the average is always zero, with the positive and negative deviations cancelling out. Instead, we can square each of the deviations and take the average of their squares. This average squared deviation is called the *variance* of the population. To make it more comprehensible, we can take its square root, and this gives the *standard deviation* (SD) which, as its name suggests, tells us a typical amount by which an observation might deviate away from the population mean.

If a mathematical form is being assumed for the shape of the distribution, the corresponding formula will contain one or more adjustable quantities which enable the formula to be adapted to the problem at hand. These are called the *parameters* of the distribution. If the values of the parameters are known, we can calculate from them the mean and standard deviation, and also the probability of observing different values of the variate—we could, in fact, draw a diagram of the distribution, just as we could draw a circle if we knew its parameters, the position of the centre and the radius. The Normal distribution, as an example, has two parameters denoted by μ and σ. It so happens that these are equal to the mean and standard deviation of the distribution. The Normal distribution is *symmetric*, with equal deviations above and below the mean having equal probabilities; as a result the median and the mode are both equal to the mean. In a Normal distribution, some two thirds of all observations lie within 1 SD above or below the mean, and 95% lie within 2 SD above or below the mean.

The important thing to emphasise at this point is that what you *know*, as researcher or reader, are the variate values in the sample. What you would like to know is the description of the population, but about this there will inevitably be an element of *uncertainty*, and statements about the population will need to be phrased in the language of probability. There are two basic ways in which standard statistical methods are commonly used to generalise from the sample values in order to make probability statements about the population. These will be the subject of the next note in this series.