STATISTICS FROM THE INSIDE

4. Sampling distributions

M J R Healy

In note 2 of this series I have described the sort of statistical reasoning that underlies significance testing. While estimation is of far more practical importance, it will have been seen in note 3 that the construction of confidence intervals is very closely tied up with the ideas of significance testing—effectively, a hypothetical value for a population parameter is inside a confidence interval if its use in a null hypothesis renders the data non-significant at an agreed level. It may be of interest, then, to describe in a bit more detail just how significance probabilities are arrived at.

To take a simplistic example, suppose we have a sample of rectal temperatures taken on 20 healthy babies and that we adopt as a null hypothesis the assertion that the mean temperature in the population from which the sample is drawn is 37.0° . The mean of the 20 sample values is (say) 37.3° , a deviation of 0.3° from the hypothetical value. What is the probability of getting a deviation as large as this?

We can approach the question by doing a thought experiment. We have a hypothetical population of temperatures with a mean of 37.0° . Imagine that we draw a very large number of samples of size 20 from this population and calculate the mean of each of the samples. By doing this, we are constructing a quite new population. Its items are samples of 20 and its variate is the mean of each sample. Our actual data provide us with a single item from this new population with a variate value of 37.3 , and what we need is the probability of obtaining a variate value as extreme as the one we have actually observed. We could get this if we were to know the probability distribution that describes our new population. This distribution is known as the *sampling distribution* of the sample mean. The original distribution of individual temperatures is called the *parent distribution*.

Statistical theory tells us several things about the new distribution. To begin with, if the parent distribution is Normal, then the sampling distribution of the mean will be Normal also. However, a remarkable result, known to statisticians by the mysterious name of the central limit theorem, states that the shape of the sampling distribution of the sample mean comes rapidly closer to that of a Normal distribution as the sample size increases, almost irrespective of the shape of the parent distribution from which the individual data values were drawn. This is one reason for the key part played by the Normal distribution in statistical practice.

Naturally enough, the parameters of the sampling distribution are related to those of the parent distribution and statistical theory provides some details. For a start, the means of the two distributions are the same, irrespective of the shape of the parent. This says that, on average over a long sequence of samples, the sample mean is equal to the population mean. In the jargon, as an estimate of the population mean, the sample mean is *unbiased*.

A second theoretical result states that the variance of the sampling distribution of the mean (the square of its standard deviation) is given by the variance of the parent distribution divided by the sample size. In other words, if the standard deviation of the parent distribution is σ and the sample size is n , the standard deviation of the sampling distribution of the mean is $\sigma/\sqrt{n}$. This in turn tells us the extent to which the mean of a sample is less variable than a single observation; it is a quantitative version of what is loosely referred to as the law of averages. For brevity, the standard deviation of the sampling distribution is usually called the *standard error* of the estimate in question.

Returning to our original question, the mean of our (hypothetical) parent distribution is 37.0° , so this is also the mean of the sampling distribution of the mean. Just for the moment, pretend that we know that the standard deviation of the parent distribution is 0.5° . Then that of the sampling distribution (the standard error of the sample mean) would be $0.5/\sqrt{20}=0.112^{\circ}$. The observed sample mean of 37.3° deviates from the mean of its sampling distribution by an amount of 0.3° , and this is equal to $0.3/0.112=2.68$ standard errors. From a table of the Normal distribution, the probability of a deviation as large as this (in either direction) is 0.0074; this is the significance probability we are seeking.

... or it would be if we had the knowledge we pretended we had about the standard deviation of the parent population. In practice we almost never have such knowledge—just as with the mean, we have to form an estimate of the standard deviation based on the values in the sample. We can postpone for the moment the details of how this is done; suppose that the result is an estimated standard deviation $s=0.48^{\circ}$, say (note the use of the italic letter s to distinguish the estimated quantity from its population counterpart σ). The estimated standard error of the mean is now $0.48/\sqrt{20}=0.107^{\circ}$ and the observed deviation amounts to $0.3/0.107=2.80$ *estimated* standard errors. This

is slightly larger than before, when we were pretending that we knew the population variance, but this may, of course, be a chance effect—it is quite possible that the estimate of the population standard deviation calculated from our sample is by chance smaller than the true value. To allow for this, we look up the probability, not in the table of the Normal distribution, but in that of the quantity t introduced by Student (the modest pseudonym of W S Gossett, head brewer at Guinness's Park Royal brewery). The probability corresponding to 2.80 standard errors in the Normal distribution is 0.0051; in the table of t , assuming a sample of size 20, it is distinctly less extreme, at 0.0113.

A very large number of statistical questions can be answered by extending the sort of argument we have used above. Essentially we start with a quantity (such as a sample mean, or the difference between two means, or a regression coefficient) which has, at least approximately, a Normal sampling distribution. We calculate the deviation of this quantity from some hypothetical value (often zero). We estimate from the data the standard deviation of the parent population and from this we derive the estimated standard error of the quantity we are using. We can then calculate t as the deviation divided by its estimated standard error and look this up in the appropriate tables.

Two subsidiary issues can be conveniently taken up here. First, how do we go about estimating the population standard deviation from the sample values? The population variance (the square of the standard deviation) is defined as the average value over the population of the squared deviations from the population mean—in symbols, the average of $(x-\mu)^2$ where x stands for a variate value and μ for the population mean. It is tempting to estimate this by taking the average over the sample of the corresponding sample quantities, $(x-m)^2$ where m stands for the sample mean. There is an oddity here, though. It can be shown quite easily that the straight average of $(x-m)^2$ is smaller than that of $(x-a)^2$ for any other value of a . In particular, we can be sure that the straight average of $(x-m)^2$ is likely to be smaller than that of $(x-\mu)^2$, even though we do not know what μ is!

This tendency to underestimate can easily be counteracted though. Instead of dividing by n , the sample size, when forming the average, we divide by $(n-1)$. This divisor is called the *degrees of freedom* of the estimated variance.

Some of the magic can be stripped from this divisor by noting that the sum of the deviations from the sample mean $(x-m)$ is *exactly* zero. This means that if you tell me $(n-1)$ of the deviations, I can tell you the last one without looking at the data. Once the sample mean has been calculated, there are only $(n-1)$ pieces of independent information left. The difference between dividing by n and $(n-1)$ is not considerable except in very small samples, but there are more complex situations where the degrees of freedom differ more markedly from the sample size and the concept is an important one. The facility provided by some calculators for dividing by n rather than by $(n-1)$ (also advocated by some textbooks when the sample size exceeds some entirely arbitrary limit) should be avoided—why bother whether your sample is 'large' or 'small' when one method is correct for both?

The second topic relates to the important distinction between the standard deviation of the parent population (the SD) and that of the sampling distribution (the standard error or SE). Both can be useful and each can be calculated from the other, but in preparing a scientific paper it is helpful to the reader if each one is quoted where it is appropriate.

The SD is a quantity that describes the variability of the parent population. It is thus usually needed in the 'subjects and methods' section of a paper. If the study involves two or more groups of patients, the reader will wish to be assured that the groups are comparable in their baseline data, and this comparability should include variability as well as mean values.

The SE on the other hand is related to the sample mean and is a measure of its precision. It is used as described above in assessing the significance of differences between the sample mean and a hypothetical value (or between two sample means). It will therefore usually be appropriate in the 'results' section of a paper and the associated tables and diagrams. The choice between quoting the SE and related confidence intervals has been discussed in my note 3.

Whether the SD or the SE is being quoted, it is extremely important to make it clear to the reader which is being used. The $\pm$ sign by itself is far too ambiguous (it may even indicate a confidence interval) and should not be used.