amc technical brief

Analytical Methods Committee No.14. Oct 2003 The Royal Society of Chemistry 2003

A glimpse into Bayesian statistics

Many analytical chemists find the logic of hypothesis tests and confidence intervals hard to follow. What looks like a probability statement about a true concentration is in fact an assertion about random intervals, involving data we did not observe but might have. There is another way. Bayesian statistics allows, indeed insists on, probabilities for hypotheses.

An example

Consider the simple example of analysing a material to test it against a specification. Suppose there is an upper limit $c_L = 10$ units for an acceptable concentration of an impurity in the material, and by analysis we obtain a single measurement $c_m = 10.7$ of the concentration in this particular sample. The analytical method is unbiased and has known precision (standard deviation) 0.4 units. Thus the variance of the measurement is $v_m = 0.4^2 = 0.16$. What is the strength of the evidence that the true concentration   in this sample exceeds the allowable limit?

A standard statistical treatment

This argues as follows. If the true value   = 10, then the measurement is drawn from a normal distribution with mean 10 and standard deviation 0.4. The probability that such a measurement is 10.7 or greater is the same as the probability that an observation from the standard normal distribution exceeds (10.7   10)/0.4 = 1.75, which is 0.04 from tables. If   < 10, this probability will be even smaller. The small probability for the observed (or more extreme) data under the hypothesis   10 is taken as evidence against the hypothesis. Either we quote 0.04 as a p-value measuring the strength of this evidence or, noting that 0.04 is less than the magic 0.05, announce that the hypothesis has been rejected at the 5% level. All this should seem fairly familiar. What may also be familiar is the common practice of interpreting the p-value as though it is the probability that the hypothesis is true. It is not. It is the probability of observing particular data given that the hypothesis is true. If we want to attach probabilities to hypotheses then we have to work in a Bayesian framework.

A Bayesian analysis

The Bayesian approach requires us to quantify our beliefs about the true value   in the form of a probability distribution. These beliefs will change when we see the result of the measurement, and the main tool in Bayesian statistics is the recipe

posterior   likelihood   prior

for updating beliefs in the light of new evidence. The workings of this formula are most easily followed in the case when   may only take one of a finite set of values, $_1$, $_2$, . . ., $_k$, and 'prior' attaches a probability to each $_i$. This prior distribution expresses our beliefs about   before observing the data. The likelihood, which also has a value for each $_i$, is the probability of observing the data given   = $_i$. Multiplying the prior probability and the likelihood for each $_i$ and then scaling so that the resulting numbers add to one over the *k* values of   gives us a new set of probabilities, the posterior distribution, which expresses our updated beliefs about . When, as in our example, it is more natural to think of   as continuous rather than discrete, prior and posterior beliefs are represented by probability density functions (pdfs), and the likelihood becomes a continuous function of , but the idea is essentially the same.

Sometimes the updating has to be done numerically, just as described above, possibly after discretising a continuous distribution. Sometimes, if the prior distribution and likelihood have compatible mathematical forms, it can be done algebraically.

Suppose that in our example our prior beliefs about   may be described by a normal distribution with mean m_p and variance v_p. This combines with the normal likelihood to give a normal posterior distribution with mean

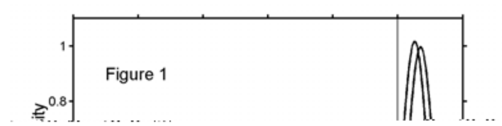
$$m = (v_m^{-1} + v_p^{-1})^{-1}(v_m^{-1}c_m + v_p^{-1}m_p) ,$$

a linear combination of the prior mean m_p and the measurement c_m with weights inversely proportional to the respective variances, and variance

$$v = (v_m^{-1} + v_p^{-1})^{-1} .$$

An 'informative prior' distribution

To get any further we need to specify the values of m_p and v_p, the prior mean and variance. If the sample of material under test comes from a manufacturing process that we have experience of, we may be able to use this experience to specify, for example, a prior mean of $m_p = 6$ and variance of $v_p = 4$. The corresponding distribution is shown in Figure 1, where it is the one cantered on 6 and spreading across the whole range.



What we are saying here is that before taking account of the measurement we are prepared to regard the material under test as a randomly chosen sample from a process that produces material with an average impurity concentration of 6 units and a spread such that about 2.5% of the material will exceed the allowable limit of 10 units.

Plugging these numbers and the values $c_m = 10.7$, $v_m = 0.16$ into the formulae above gives us a mean of