AMC Datasets—a resource for analytical scientists

Anal tical Methods Committee, AMCTB No. 72

Have you noticed, when downloading a Technical Brief from the AMC's web pages (www.rsc.org/amc), that there is a section called *AMC Datasets* listed in the leftmost column? This content was inaugurated some years ago to provide a permanent collection of interesting datasets related to analytical chemistry and its applications. The basic idea was to provide analytical chemists with material that could be used to support teaching, learning and research in statistics and chemometrics. New ideas in these fields could be tested on real and well-characterised datasets, and compared with results of other workers.

The datasets were collected from a range of activities in chemical measurement, from simple calibrations and method comparisons, through homogeneity tests, to datasets that had been used for pattern recognition or multivariate calibration. Teachers could use these as examples to demonstrate possible approaches to analysing the data, and leave a commentary on the behaviour of various mathematical approaches for future reference. Students trying an unfamiliar statistics package or an alternative statistical procedure could compare their outcome with existing commentaries from (hopefully) authoritative sources. Some interesting examples are featured below.

Calibration for aflatoxin M1 (Dataset No. 1)

The data file is shown in Box 1. (All data files show the same style of background information.) In this instance there are four repeat observations of response at each of six concentrations of the analyte. The object of such an elaborate design would be to test the calibration for curvature. The calibration plot (Fig. 1) shows no visible sign of either non-zero intercept or deviation from a straight line. The correlation coefficient is 0.9997. However, the repeat responses at each concentration provide scope for using the pure error test for linearity.

Weighted linear regression showed an intercept not significantly different from zero, but the pure error test gave a significant result (

Proficiency test results: polyunsaturated fatty acids (PUFA) in a cooking oil (Dataset No. 22)

The dataset comprises results obtained by 42 participant laboratories. The statistical procedure illustrates testing the suspicion that the distribution is bimodal, as suggested by a dotplot (Fig. 3). The method involves kernel density estimation, that is, smoothing the density of the data along the measurement axis by plausible degrees and noting the formation of modes and shoulders. (This is a type of one-dimensional unsupervised pattern recognition.)

Fig. 4 shows the outcome with smoothing parameters of 0.2 and 0.4. These values are set somewhat smaller than the reproducibility standard deviations expected (0.6) and found (0.8, robust) for this analysis, so as to detect signs of multimodality but smooth over most chance outcomes. Both graphs show visual signs of bimodality. Unfortunately, it is not possible by statistics to attach a probability to the inference of bimodality. In this instance, however, there was strong supporting evidence that two different calibration strategies (one incorrect) had been used among the participant laboratories. One involved using an internal standard, the other simply normalising the total areas under the peaks for the various fatty acids in the chromatogram. The ratio of the modal values found in

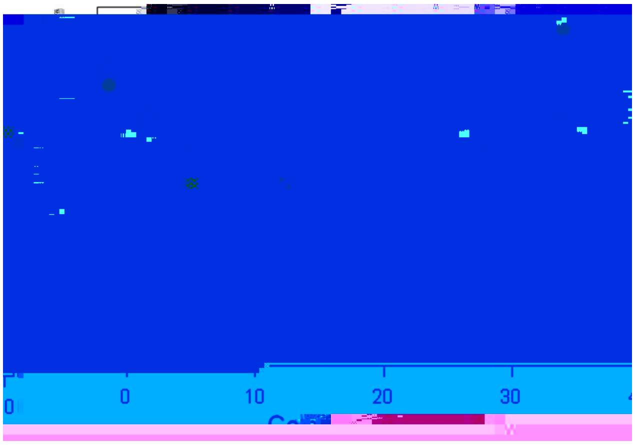

Fig. 2 A

the kernel density was very close to that expected from a consideration of the two calibration strategies.

“Homogeneity test” on a rock powder (Dataset No. 16)

discriminating criterion between a chosen subset and the disjoint subset (that is, all of the remaining objects) is the Euclidian distance of the objects from the model subspace. For present purposes separate models of two subsets were constructed and the calculated distances plotted against each other (Fig. 6). Both models provide a complete separation between the target type and all of the other types.

Feedback

If you have any observations about any of the datasets, you can post them on MyRSC (http://my.rsc.org/home) in the Group "Analytical Methods Committee—Announcements and Discussions".

B ft A C 28/07/15.

CPD Certification I certify that I have studied this document as a contribution to continuing professional development.

Name

Signature Date

Name of supervisor