amc technical brief

Editor M Thompson Analytical Methods Committee AMCTB No 4 Revised March 2006

Representing data distributions with kernel density estimates

Histograms are the usual vehicle for representing medium sized data distributions graphically, but they suffer from several defects. The kernel density estimate is an alternative computer-intensive method, which involves smoothing the data while retaining the overall structure. It is a good method of reconstructing an unknown population from a random sample of data, overcomes the problems of histograms and has many applications in analytical chemistry. An Excel add-in and Minitab macro for calculating kernel density estimates are available in AMC Software [1].

Problems with the histogram

The graphical representation of a data set is an indispensable aid to interpretation. Graphical displays facilitate visual judgements about central tendency, confidence intervals, significant difference etc. Such judgements are a valuable prelude to the use of statistics: they act as a cross-check of the statistical results, and they permit decisions about whether the distribution of the data conforms to the assumptions underlying the theory of the statistical test. The tools most frequently used by analytical scientists to visualise the distribution of univariate data are the dotplot and, for larger datasets, the histogram.

The histogram is simple to construct and provides an impression of the density distribution of the data if an appropriate choice of classes is used. If the data are a random selection, the histogram is an estimate of the population density distribution. However, the visual impression gained from a histogram can depend to an unwelcome extent on the intervals selected for the classes (*i.e.*, the number and midpoint of the bins). A reconstruction of the population density more consistent than the histogram would therefore be welcome. Computer power can now fulfil this requirement with the kernel density estimate [2, 3].

The kernel density estimate

The simple idea underlying the kernel estimate is that each data point x_i, $i=1,...,n$ is replaced by a specified distribution (typically normal), centred on the point and with a standard deviation designated by *h*. The normal distributions are added together and the resulting distribution, scaled to have a unit area, is a smooth curve, the kernel density estimate, given by

$$\hat{f}(x,h)=\frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{x-x_i}{h}\right),$$

where $\hat{f}(x,h)$ is the height of the curve at *x* (a point on the *x*-axis), and $\phi(.)$ is the standard normal density. The appearance of the kernel density, in particular the number of modes, depends critically on the value of the smoothing parameter *h*, as illustrated in Figures 1 and 2.

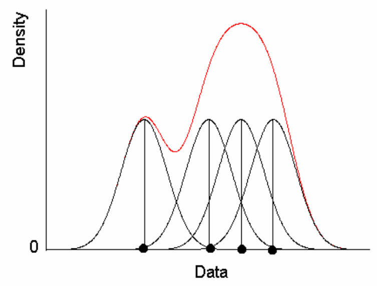

Figure 1. A normal kernel density (red line) derived from four data points (solid circles). The smoothing parameter *h* is the standard deviation of the normal kernels (black line curves).

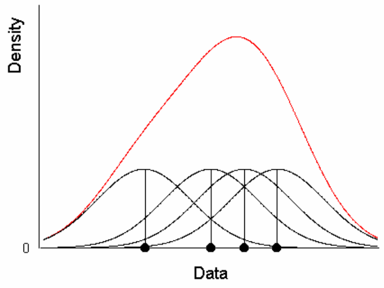

Figure 2. A normal kernel density (red line) derived from the four points shown in Figure 1, but with a value of *h* twice that shown in Figure 1.

The kernel estimate, when calculated with an appropriate value of *h*, gives a good estimate of the population density function without making any assumptions, for example, that it is a normal distribution. This is useful in examples from analytical science, where deviation from normality is common. The calculations can be programmed readily and produced as a graphic. The only complication is determining an appropriate value for *h*. This choice is context-specific and requires experience and judgement.

Examples

Here we show three examples of kernel distributions of data from interlaboratory exercises in analytical science, namely, proficiency test results from the Food Analysis Performance Assessment Scheme (FAPAS™) [4].

Figure 1 shows results obtained for the mycotoxin aflatoxin M1 in milk (FAPAS 0472). The data points alone (Figure 3) suggest the possibility of a multimodal dataset. This appearance often happens by chance in dotplots and histograms of small random samples from unimodal populations. However, comparable interlaboratory studies [5] tell us that in this instance a reproducibility standard deviation of about 14 parts per trillion should be expected. We can use this value to obtain a suitable *h* value: 0.75 times the expected value should be wide enough to smooth out any artifactual modes, but small enough to avoid undue ‘smearing’ of the data. When we construct a kernel density on this basis, we see a unimodal and almost symmetric curve (Figure 3). Close inspection shows that the curve has slightly heavier tails than a normal distribution.

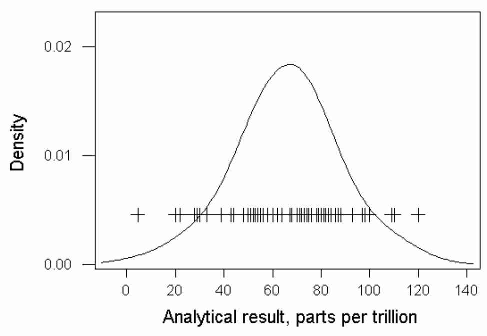

Figure 3. Analytical results for aflatoxin M1 in milk (FAPAS 0472), showing data points (crosses) and the kernel density representation (line).

Figure 4 shows results for polyunsaturated fatty acids in cooking oil (FAPAS 1416). Again the dotplot suggests that the data might be multimodal. Considerations similar to the above show that an *h*-value of 0.55 % would be appropriate, and this gave rise to a kernel density with a mode at about 39.3 and a pronounced shoulder at 40.6 %. Further investigation showed that these features were accounted for by the use of two different calibration protocols among the participants.

Figure 5 shows results for tin in fruit juice (FAPAS 0760). Here the dotplot rather strongly suggests that the data are multimodal. When we construct a kernel density, by using an *h*-value of 10 ppm, we see a curve with thee modes and a high shoulder. The most prominent mode corresponds closely with the concentration of tin spiked into the fruit juice, and presumably represents laboratories using appropriate analytical methods. The lower modes presumably represent low recovery of tin (a well-known circumstance).

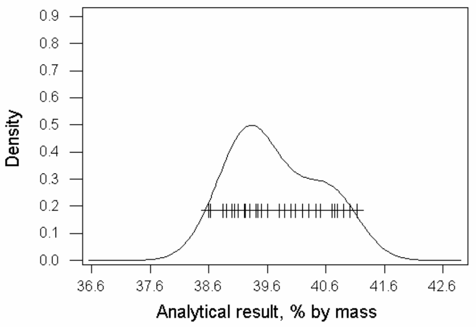

Figure 4. Analytical results for polyunsaturated fatty acids in cooking oil (FAPAS 1416), showing data points (crosses) and the kernel density representation (line).

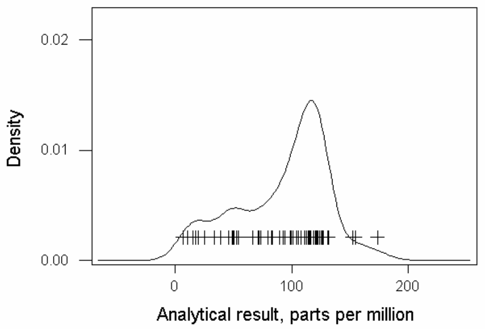

Figure 5. Analytical results for tin in fruit juice (FAPAS 0760), showing the data points (crosses) and the kernel density representation (line).

Conclusions

The examples show that the kernel density estimator is a useful method of representing the overall structure of the data. Some expertise and judgement is required for the selection of an appropriate value of the smoothing parameter *h*.

References

1. www.rsc.org/amc/
2. B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, Great Britain, 1986.
3. M.P. Wand and M.C. Jones, *Kernel Smoothing*, Chapman and Hall, London, Great Britain, 1995.
4. www.fapas.com/index.cfm
5. *Analyst*, 2000, **125**, 385-386.

This Technical Brief was produced for the Analytical Methods Committee by the Statistical Subcommittee (Chair M Thompson)

Other AMC Products (including the software mentioned in this Technical Brief can be found on: **www.rsc.org/amc/**