

amc technical brief

Analytical Methods Committee

No. 4. Jan 2001

© Royal Society of Chemistry 2001

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 and overcomes the problems of histograms. Kernel estimation is shown here for examples in analytical chemistry. A MINITAB macro for calculating kernel density estimates is available in AMC Software.

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.

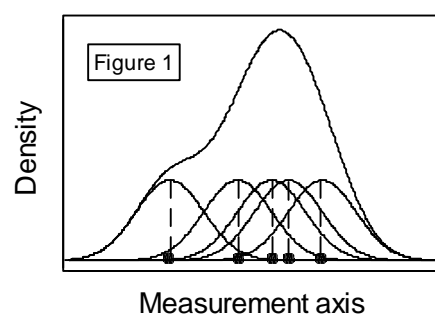
The histogram is simple to construct and provides an impression of the density distribution of the data if an appropriate choice of classes is used. If the data are a random selection, the histogram is an estimate of the population density distribution. However, the visual impression gained from a histogram 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.^{1,2}

The kernel density estimate

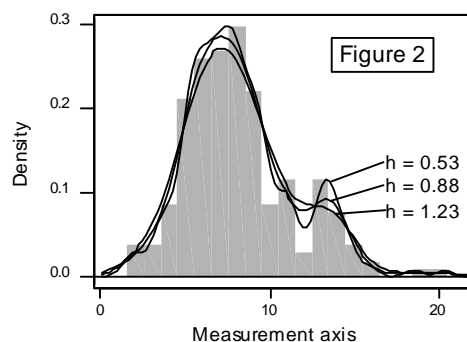
The simple idea underlying the kernel estimate is that each data point is replaced by a specified distribution (typically normal), centred on the point and with a standard deviation designated by h . The normal distributions are added together and the resulting distribution, scaled to have a unit area,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

An algorithm for the calculation is shown in Appendix 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 in the macro language of a statistics package and produced as a graphic.



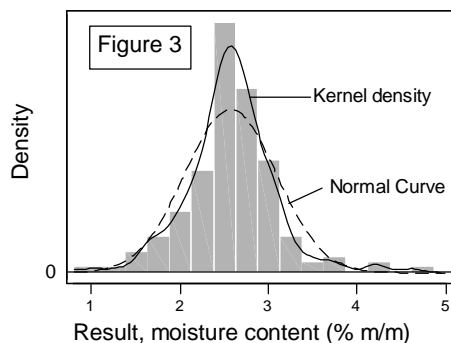
The only complication lies in estimating the appropriate value of h , which controls the degree of smoothing (Fig 2).



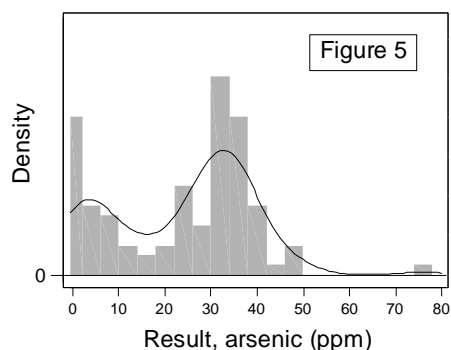
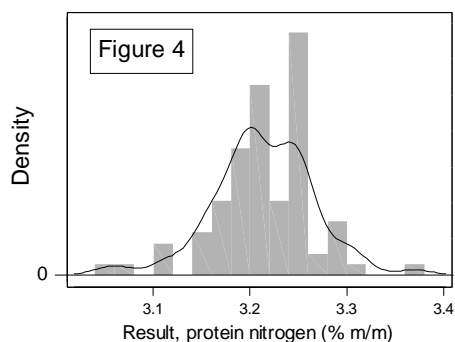
Examples

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 (FAPASTM).³ The plots were produced with automatic selection of h by using the MINITAB macro.⁴

The first example, illustrating data where there is no suggestion of bimodality, shows results of the measurement of moisture content in a meat product (Fig. 3). Here the distribution is unimodal and approximately symmetrical but somewhat 'peaky' and with heavy tails. The kernel density represents the data better than a normal distribution based on the sample statistics.



A second example, with possible evidence of bimodality, is shown in Fig. 4. The data are results of the determination of the protein nitrogen content in a meat product. Here there is a suggestion of more than one mode. In a further example, clear evidence of bimodality can be seen in the results of the determination of arsenic in a crabmeat material (Fig. 5). In that instance there are two distinct populations.



Conclusions

The examples show that the kernel density estimator is a useful method of representing the overall structure of the data. The automatic choice of h , using the method described in Appendix 2, is likely to provide a more reliable result than a histogram using a default bin width or one chosen subjectively.

References

1. B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, Great Britain, 1986.
2. M.P. Wand and M.C. Jones, *Kernel Smoothing*, Chapman and Hall, London, Great Britain, 1995.
3. www.csl.gov.uk
4. www.rsc.org/lap/rsccom/amc

This work was prepared by Philip J Lowthian and Michael Thompson, School of Biological and Chemical Sciences, Birkbeck College (University of London), Gordon House, 29, Gordon Square, London, UK WC1H 0PP, under the SAC Fellowship arrangements with financial support from the Analytical Methods Trust and the Ministry of Agriculture, Fisheries and Food (now from the Food Standards Agency).

Appendix. Algorithm to calculate kernel densities

- Determine the x -axis range for plotting, normally from minimum x to maximum x .
- Calculate sample standard deviation, upper quartile, lower quartile and hence the inter-quartile range (IQR) of the data.
- Calculate $h_{opt} = 0.9 * \min(\text{standard deviation}, \text{IQR}/1.34) * n^{-1/5}$.
- (This is a quick and straightforward method of calculating the optimum h value, which produces a satisfactory window width or an initial value that can be adapted by the user.)
- Calculate the kernel estimator for each selected point on the x -axis. The curve $\hat{f}(x;h)$ is obtained from the following formula:

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

where x is the selected point on the x -axis, x_i is a data point from x_1, \dots, x_n , $f(\cdot)$ is the standard normal density, and h is the selected value.

Test data sets and their kernel densities can be found on the AMC Website.⁴

AMC Technical Briefs (Editor M Thompson) are informal but authoritative bulletins on technical matters of interest to the analytical community. They are prepared by the Analytical Methods Committee of the Analytical Division of the RSC, and are carefully scrutinised for accuracy. Correspondence should be addressed to: The Secretary, The Analytical Methods Committee, The Royal Society of Chemistry, Burlington House, Piccadilly, London W1V 0BN. Other AMC Technical Briefs can be found on:

www.rsc.org/lap/rsccom/amc/amc_index.htm