Optimised Scoring in Proficiency Tests

Michael Thompson
School of Biological and Chemical Sciences
Birkbeck College (University of London)
Malet Street
London WC1E 7HX
m.thompson@bbk.ac.uk

Criteria for an ideal scoring method

- Adds value to raw results.
- Easily understandable, no arbitrary scaling transformation.
- Is transferable between different concentrations, analytes, matrices, and measurement principles.

The z-score

Result

“Assigned value”

Scheme provider’s best estimate of true value

$$z \qquad \frac{x \qquad x_A}{p}$$

“Target value” or

“standard deviation for proficiency”

Determining an assigned value

- Reference laboratory result
- Certified reference material(s)
- Formulation
- Consensus of participants’ results

"Health warnings" about the consensus

- The consensus is not necessariledseoicslbout

What exactly *is* a ‘consensus’?

- Mean? -

Finding a 'consensus'
—the tools of the trade

- Robust mean and standard deviation
- Kernel density mode and its standard error
- Mixture model representation

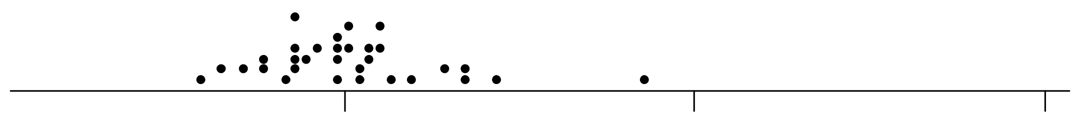
Robust mean and standard deviation

$$\hat{\ }_{rob},\quad \hat{\ }_{rob}$$

- Robust statistics is applicable to datasets that look like normally distributed samples contaminated with outliers and stragglers (*i.e.*, unimodal and roughly symmetric).
- The method downweights the otherwise large influence of outliers and stragglers on the estimates.
- It models the central 'reliable' part of the dataset.
- The estimates are found by a procedure, not a formula.

$\mathbf{x}^{\mathbf{T}}$ x_1 x_2 x_n

When can I safely use robust estimates?



The robust mean as consensus

- The robust mean provides a useful consensus in the great majority of instances.

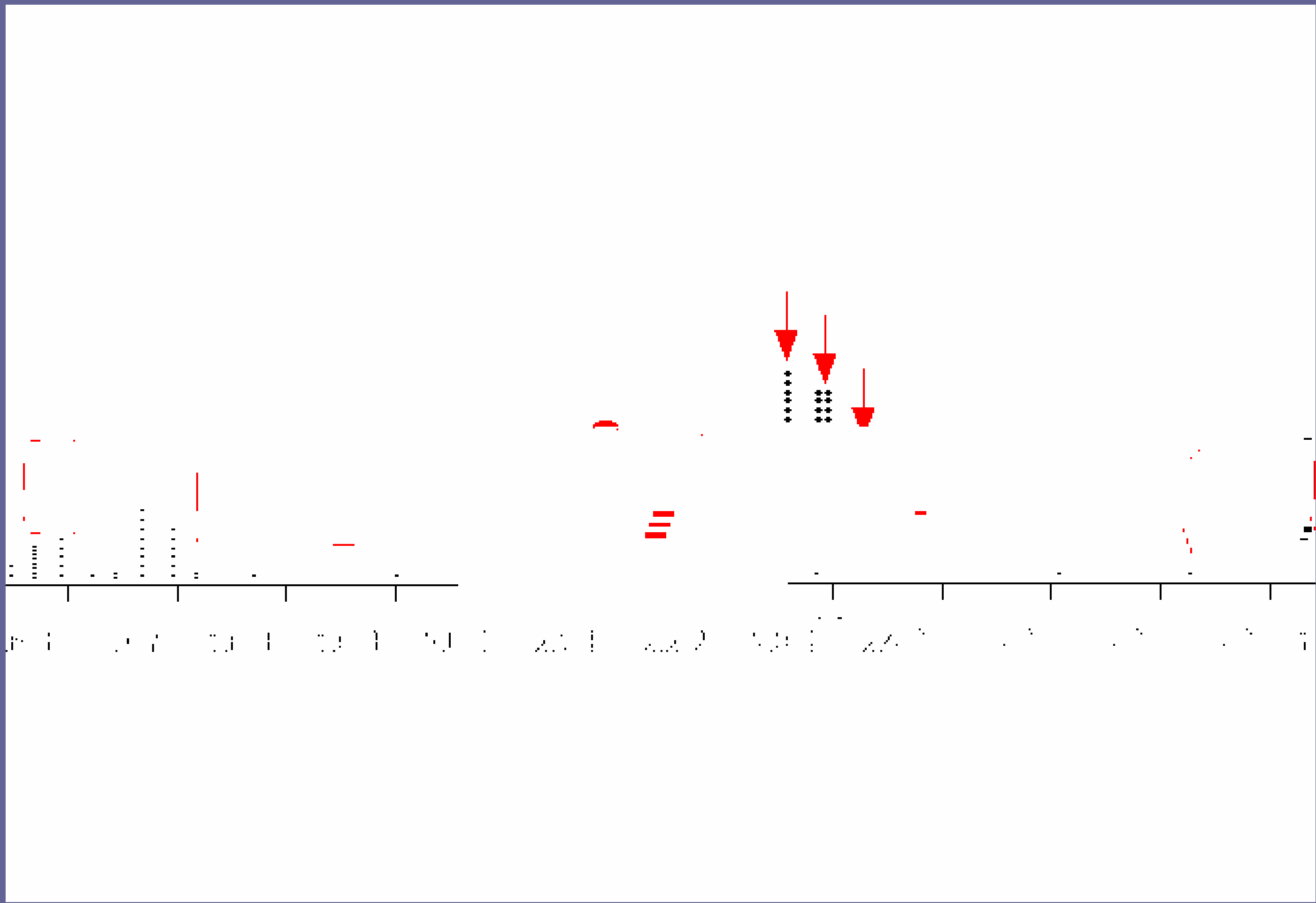
- The uncertainty of this consensus can be safely taken as $u \; x_a \qquad \hat{\ }_{rob} / \sqrt{n}$

Finding a 'consensus' —the tools of the trade

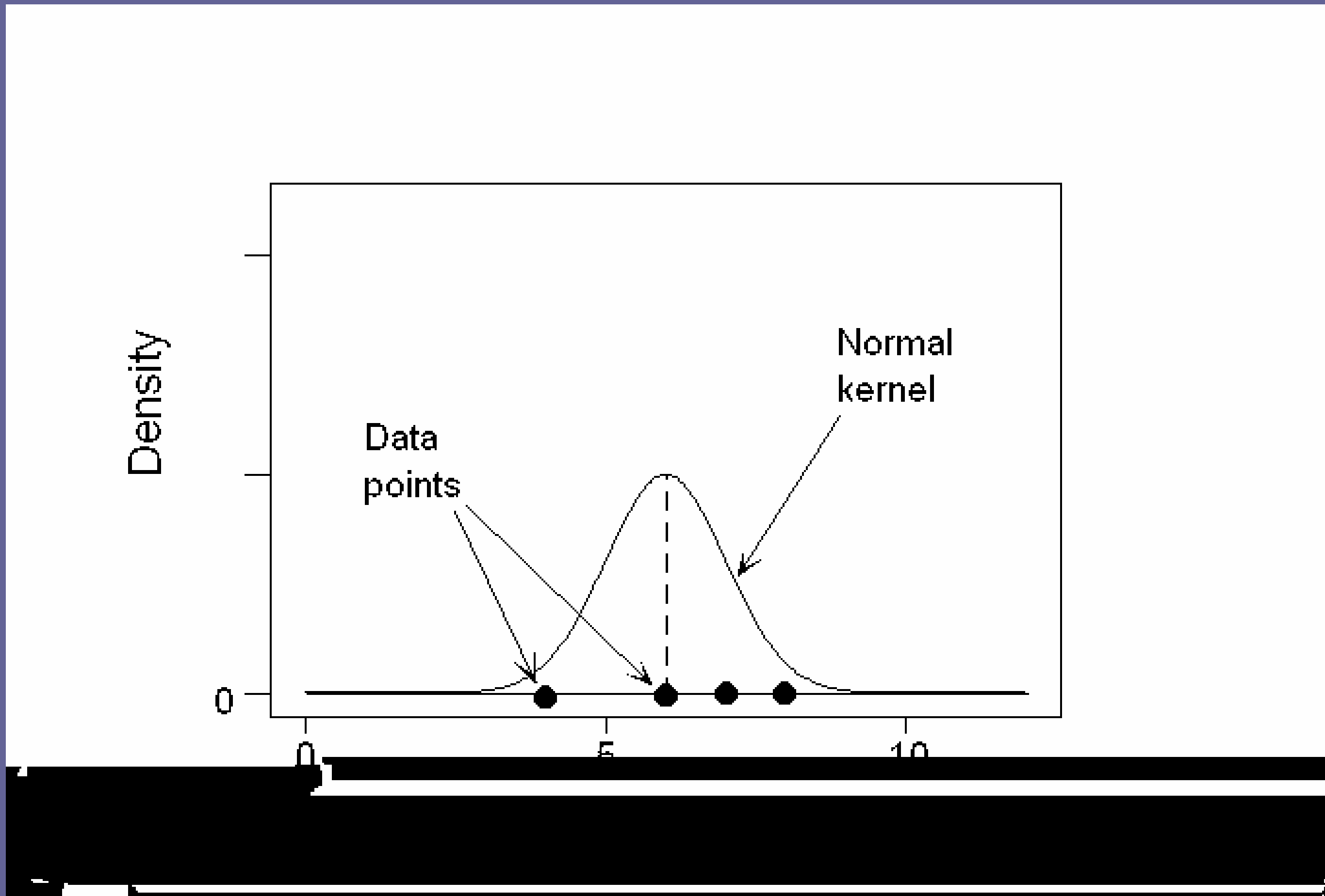
- Robust mean and standard deviation
- Kernel density mode and its standard error
- Mixture model representation

The mode as a consensus

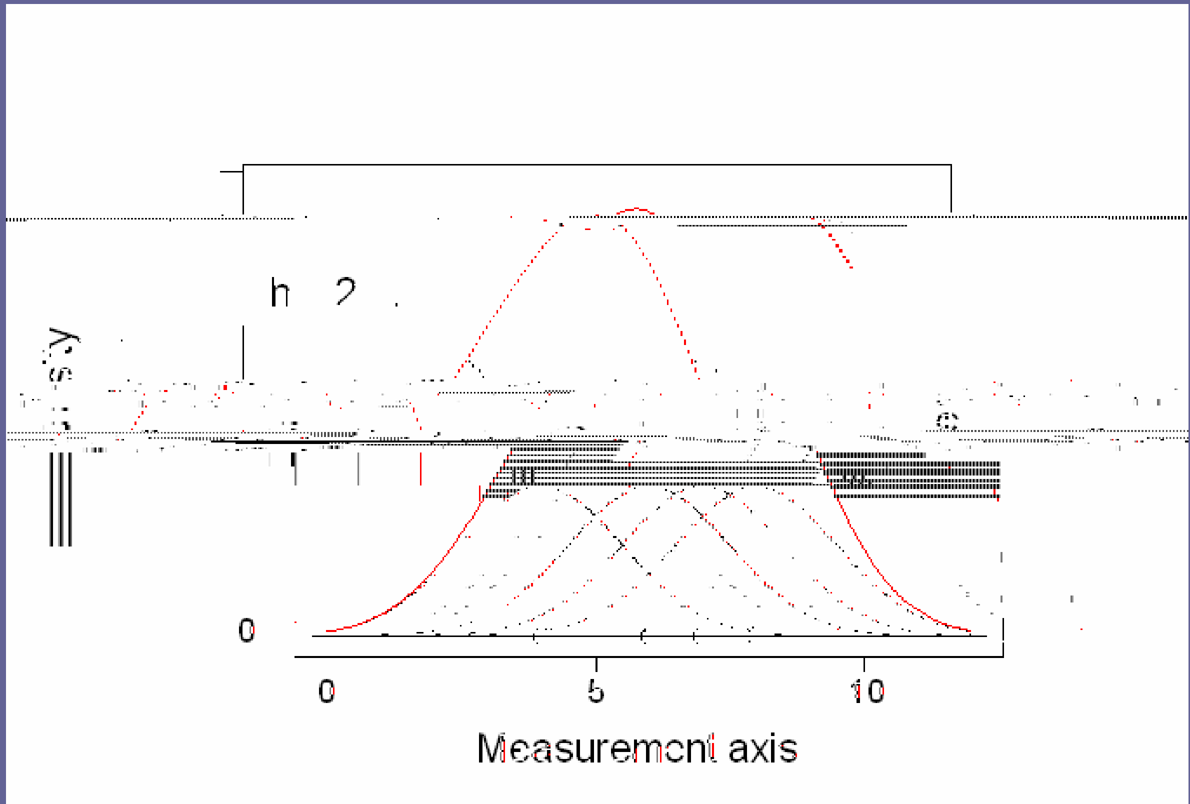
Can I use the mode? How many modes? Where are they?



A normal kernel

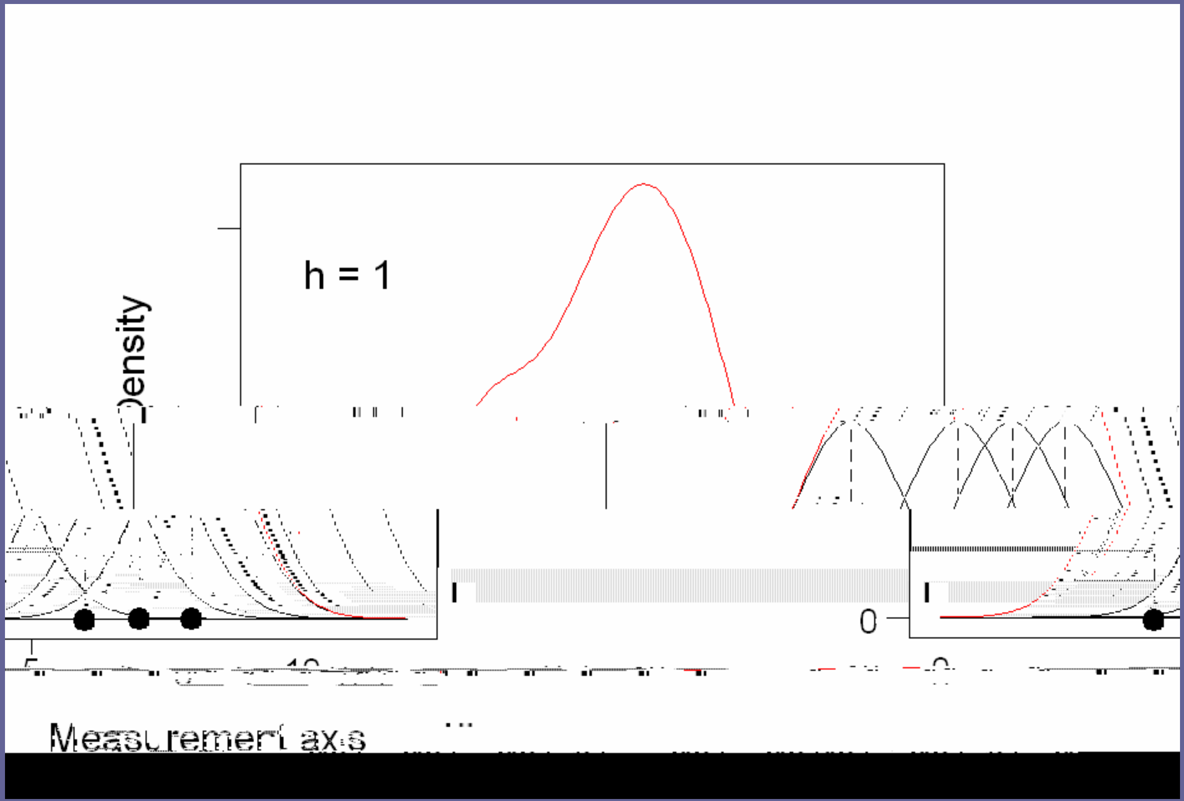


A kernel density



Reference: *AMC Technical Brief* No. 4. (www.rsc.org/amc)

Another kernel density: same data, different *h*



Reference: *AMC Technical Brief* No. 4. (www.rsc.org/amc)

Uncertainty of the mode

- The uncertainty of the consensus can be estimated as the standard error of the mode by applying the bootstrap to the procedure.

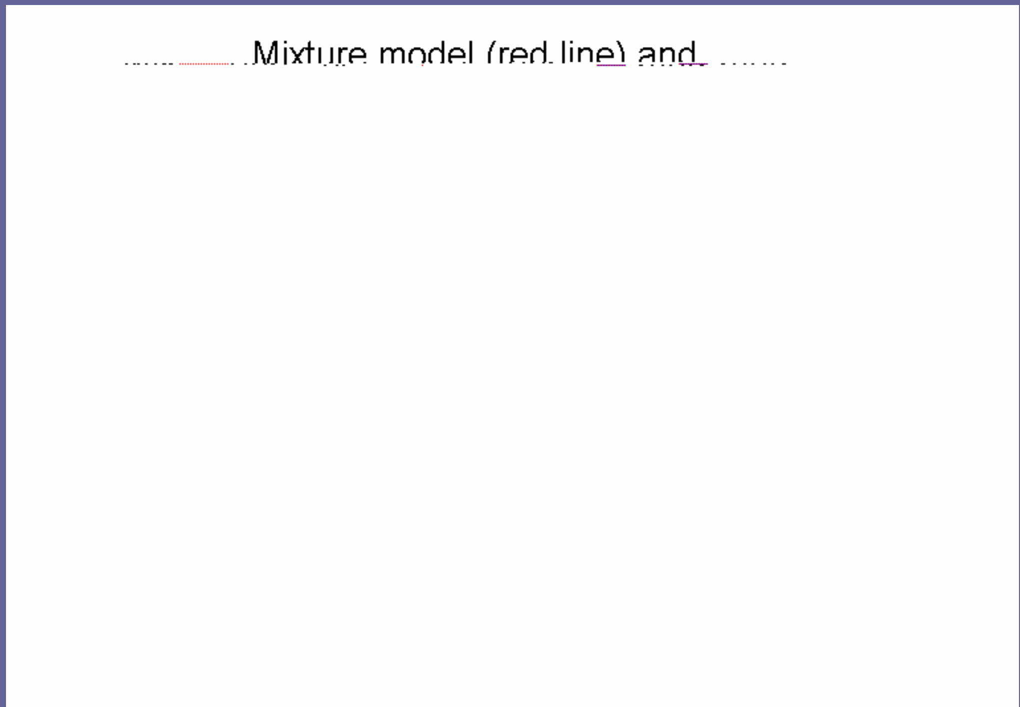
- The bootstrap is a general procedure, based on resampling, for estimating standard errors of complex statistics.

- Reference: *Bump-hunting for the proficiency tester – searching for multimodality.* P J Lowthian and M Thompson, *Analyst*, 2002,**127**, 1359-1364.

Finding a ‘consensus’ —the tools of the trade

- Robust mean and standard deviation
- Kernel density mode and its standard error
- Mixture model representation

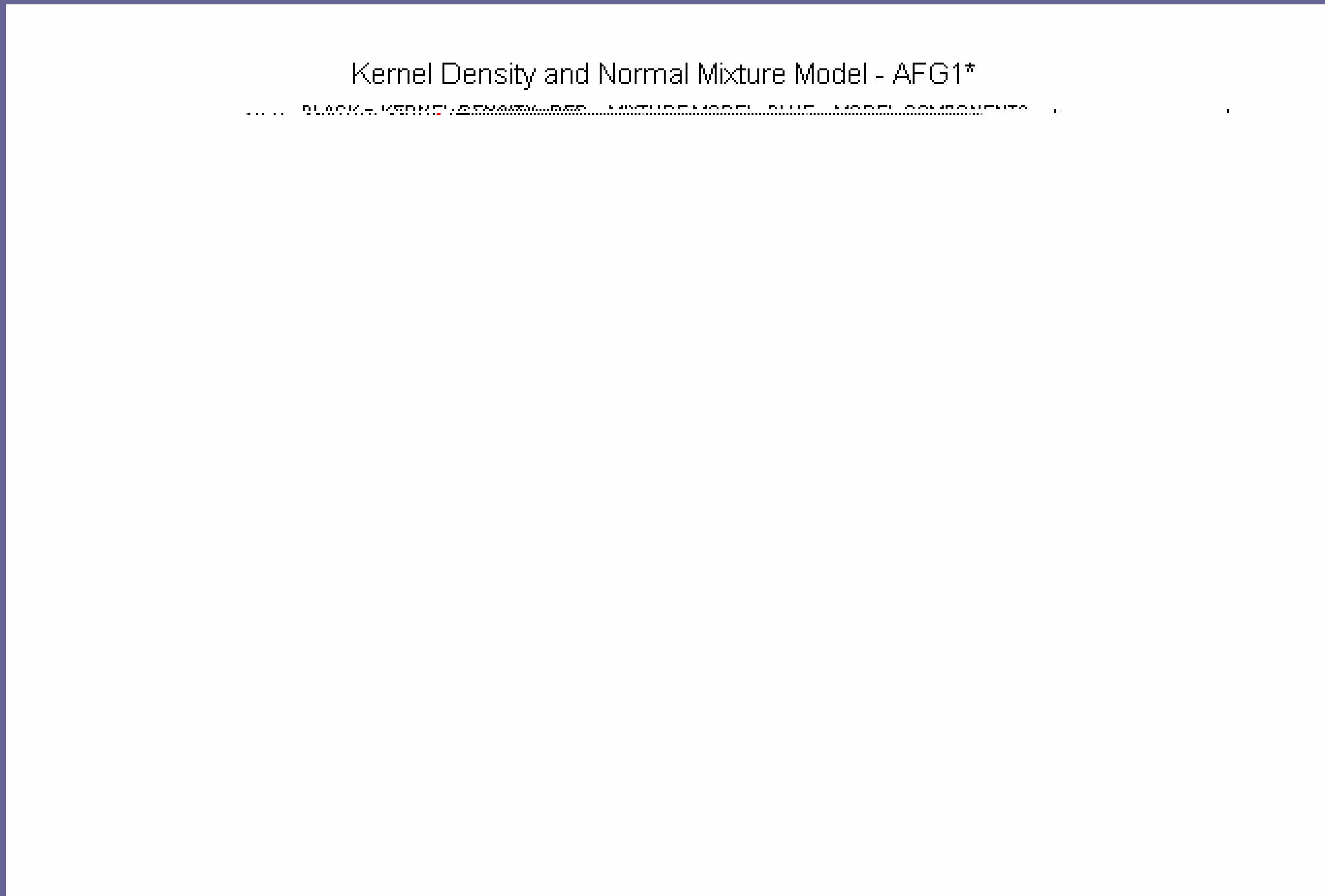
Mixture models and consensus



- For each component you can calculate:

  - a mean
  - a variance
  - a proportion

2-component normal mixture model and kernel density



The normal mixture model

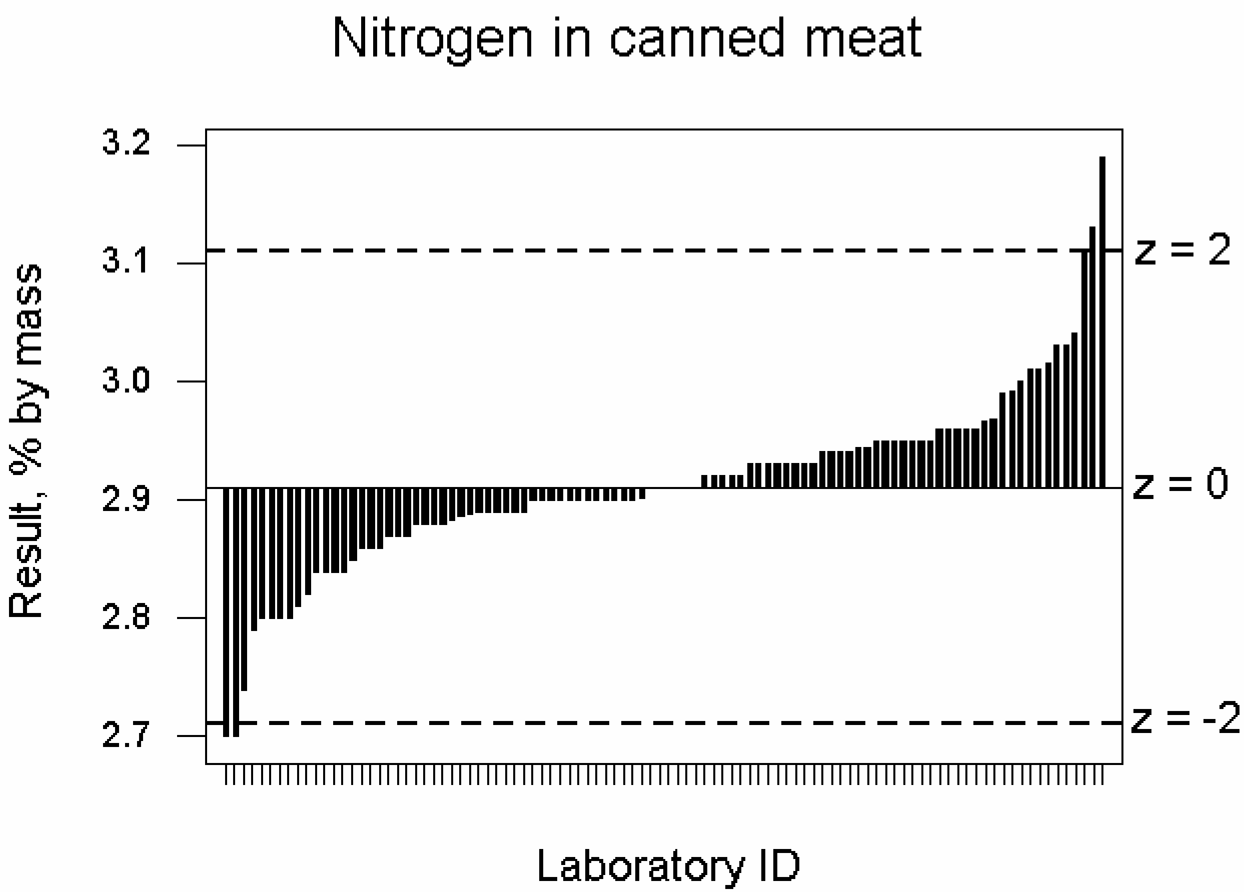
$$f(y) \quad {}_{j\quad 1}^{m} p_j f_j(y), \quad {}_{j\quad 1}^{m} p_j \quad 1$$

$$f_j(y) \quad \frac{\exp(\quad (y \qquad {}_j)^2 / 2 \quad {}^2)}{\sqrt{2}}$$

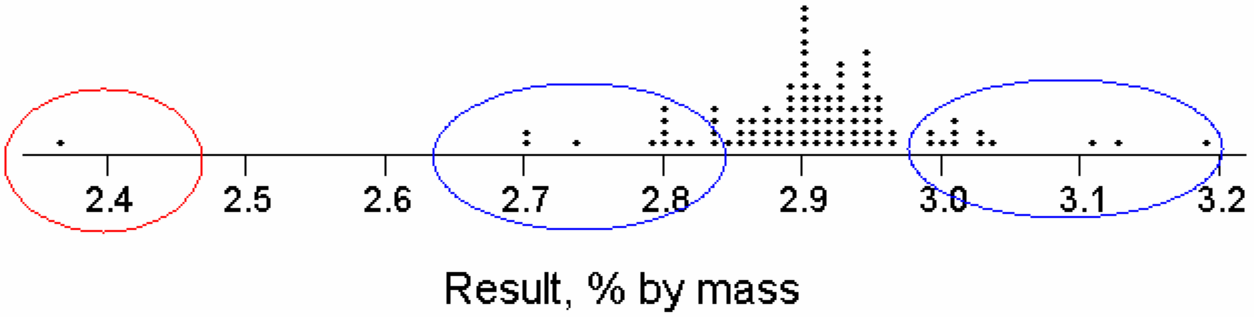
References: *AMC Technical Brief* No 23, and *AMC Software.*
Thompson*, Acc Qual Assur,* 2006, **10**, 501-505.

Example datasets

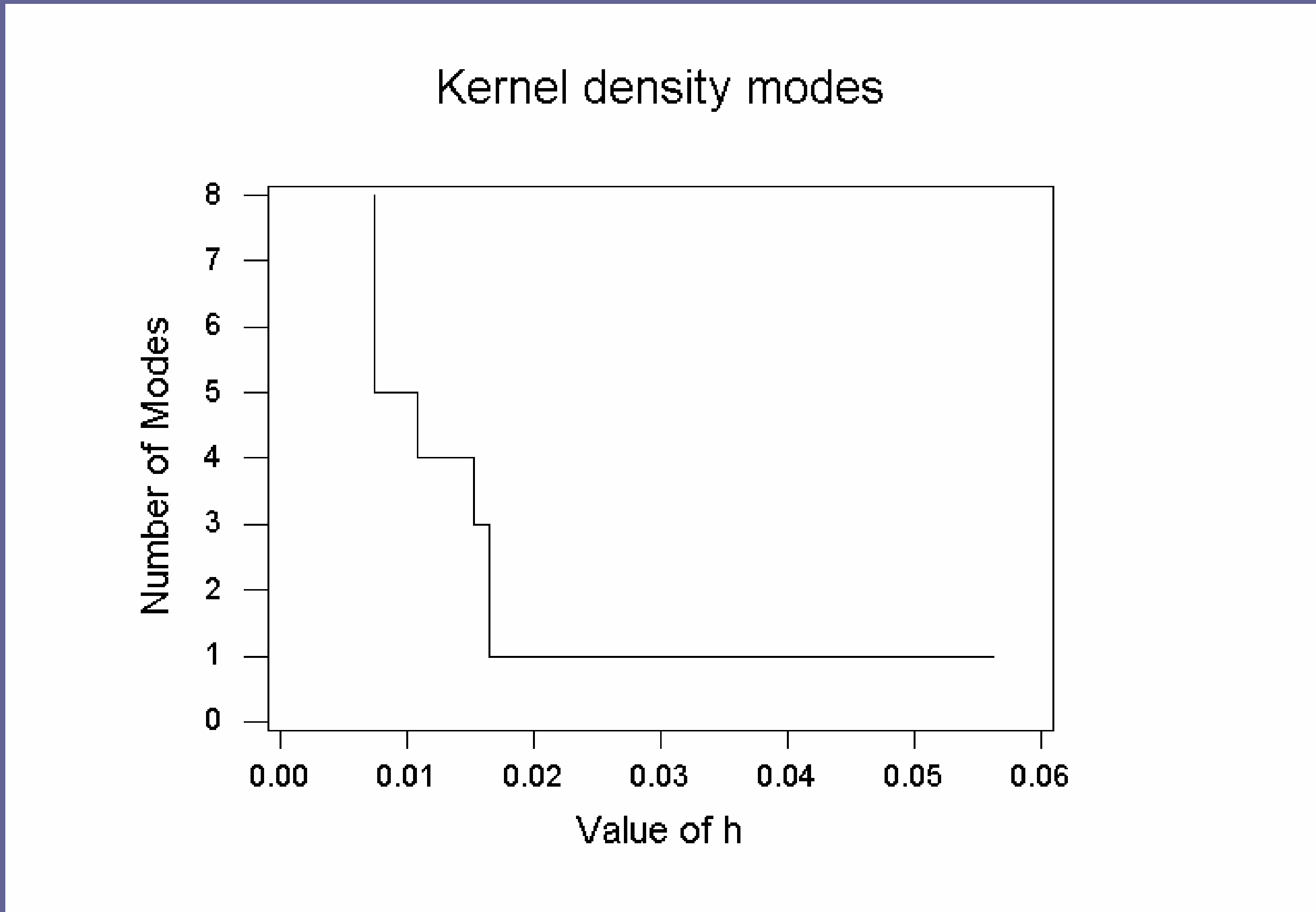
Example dataset 1



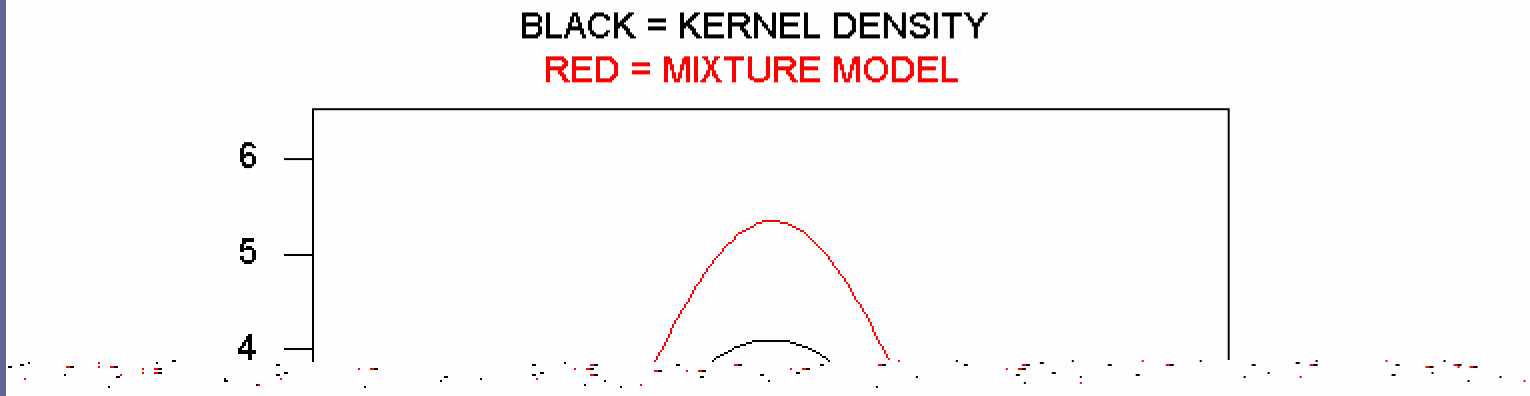
Nitrogen in canned meat

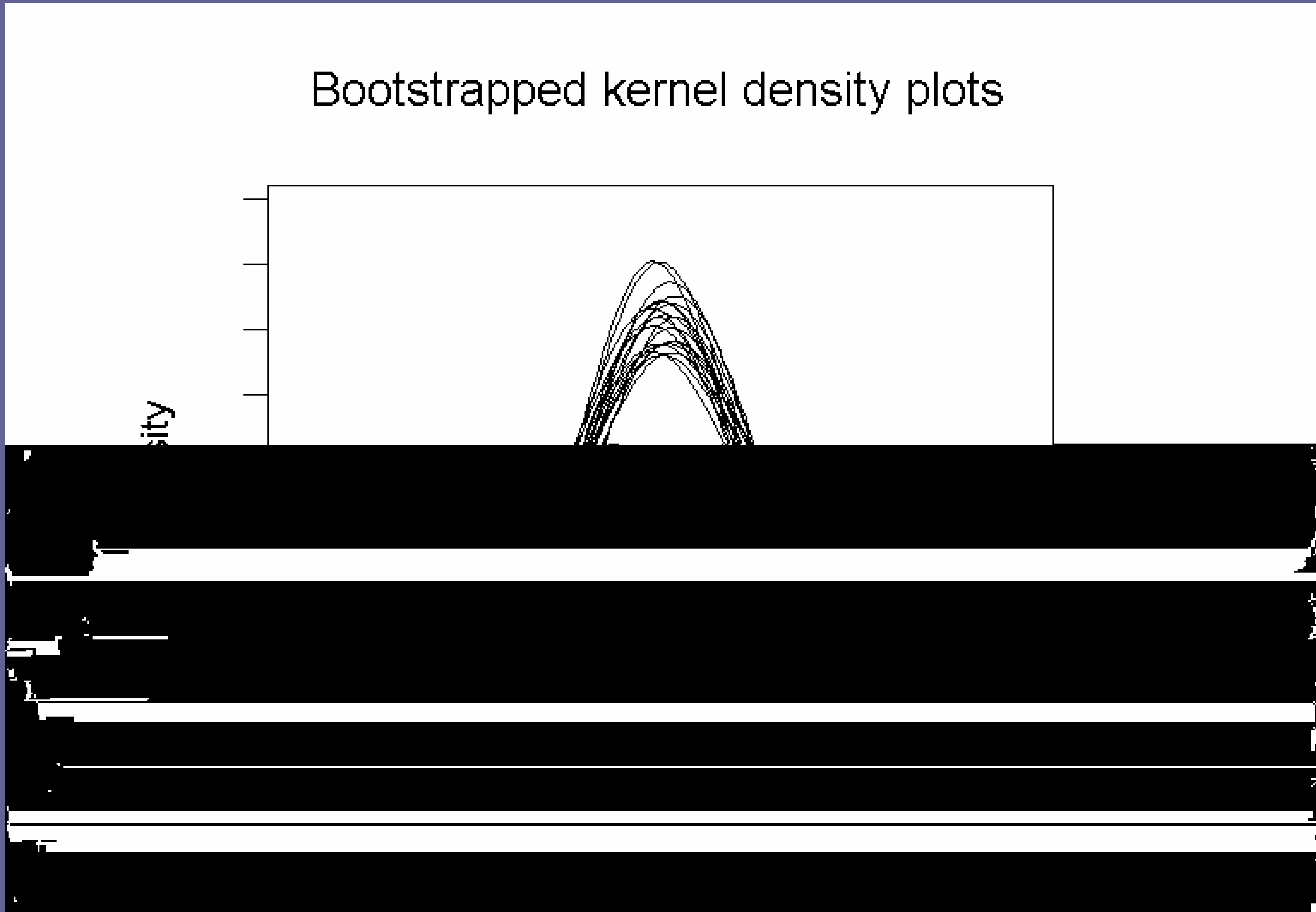


Number of modes vs smoothing factor *h*



Nitrogen in canned meat




Bootstrapped kernel density plots

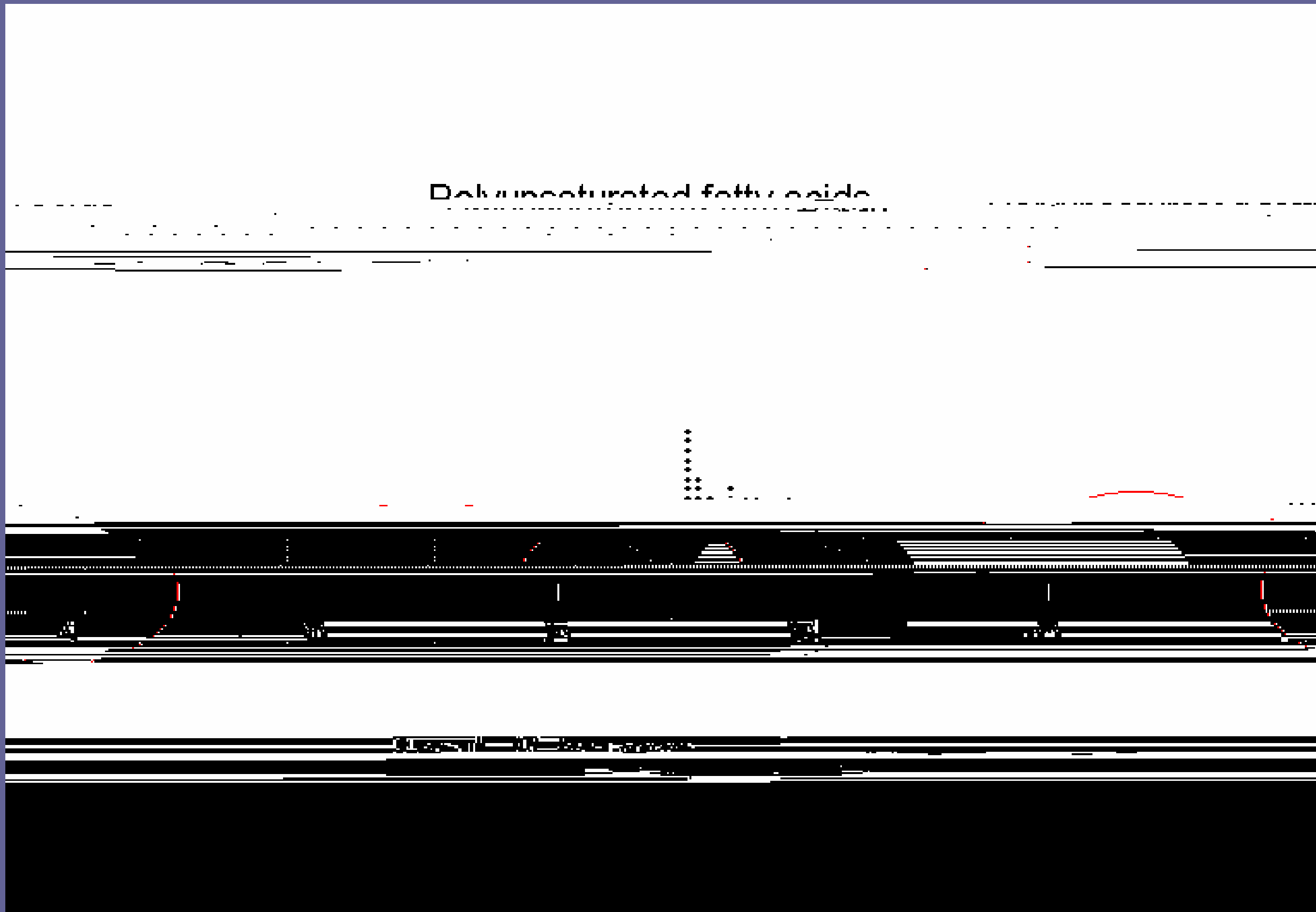
Statistics: dataset 1

	^	^	*se* ^
Robust	2.912	0.056	0.0056
Kernel density mode	2.912	-	0.0056
Mixture model	2.913	0.075	0.0075

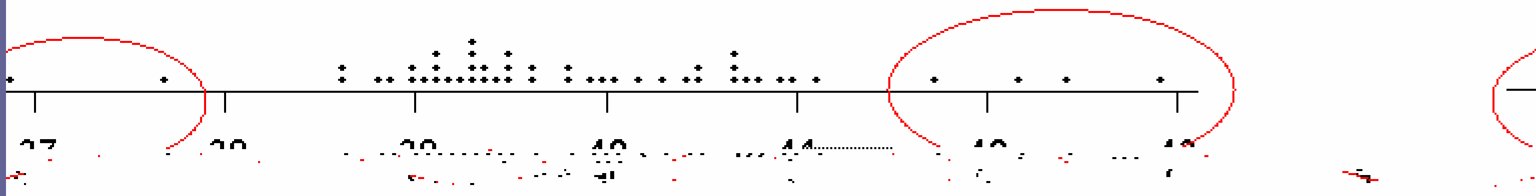
Skewed/multimodal distributions

- Skews and extra modes can arise when the participants' results come from two or more inconsistent methods.
- Skews can also arise as an artefact at low concentrations of analyte as a result of common data recording practices.
- Rarely, skews can arise when the distribution is truly lognormal (*e.g.*, in GMO determinations).

Example dataset 2

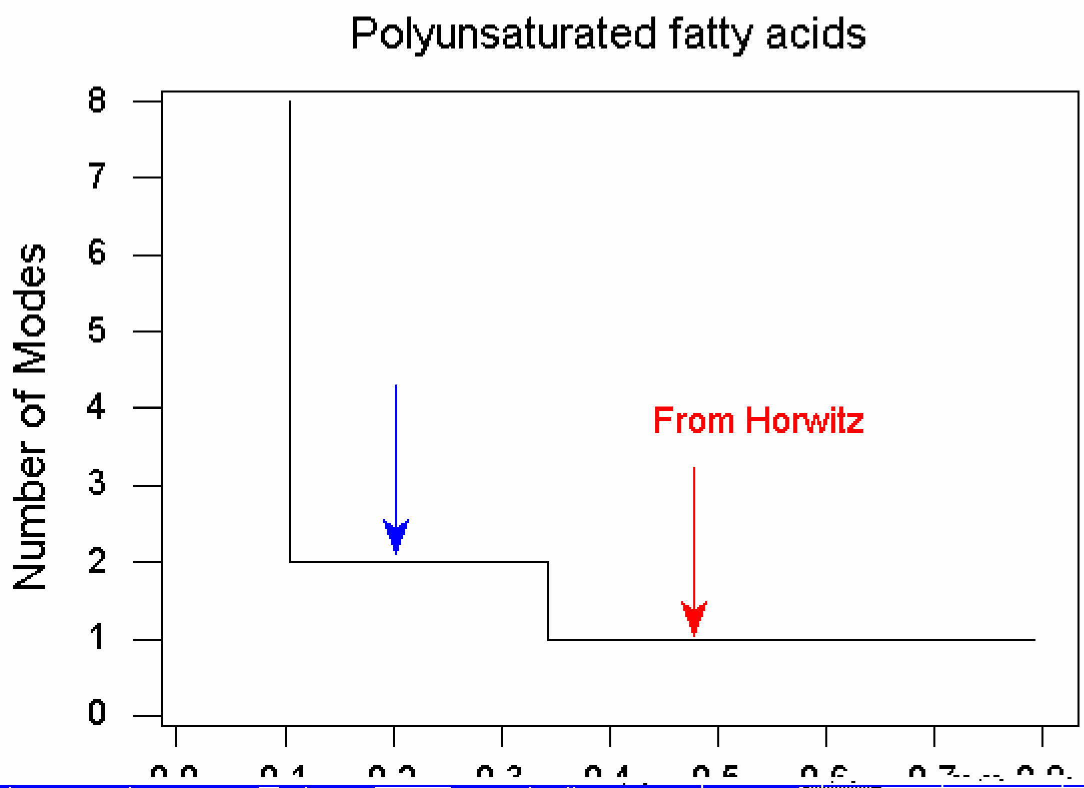


Polyunsaturated fatty acids

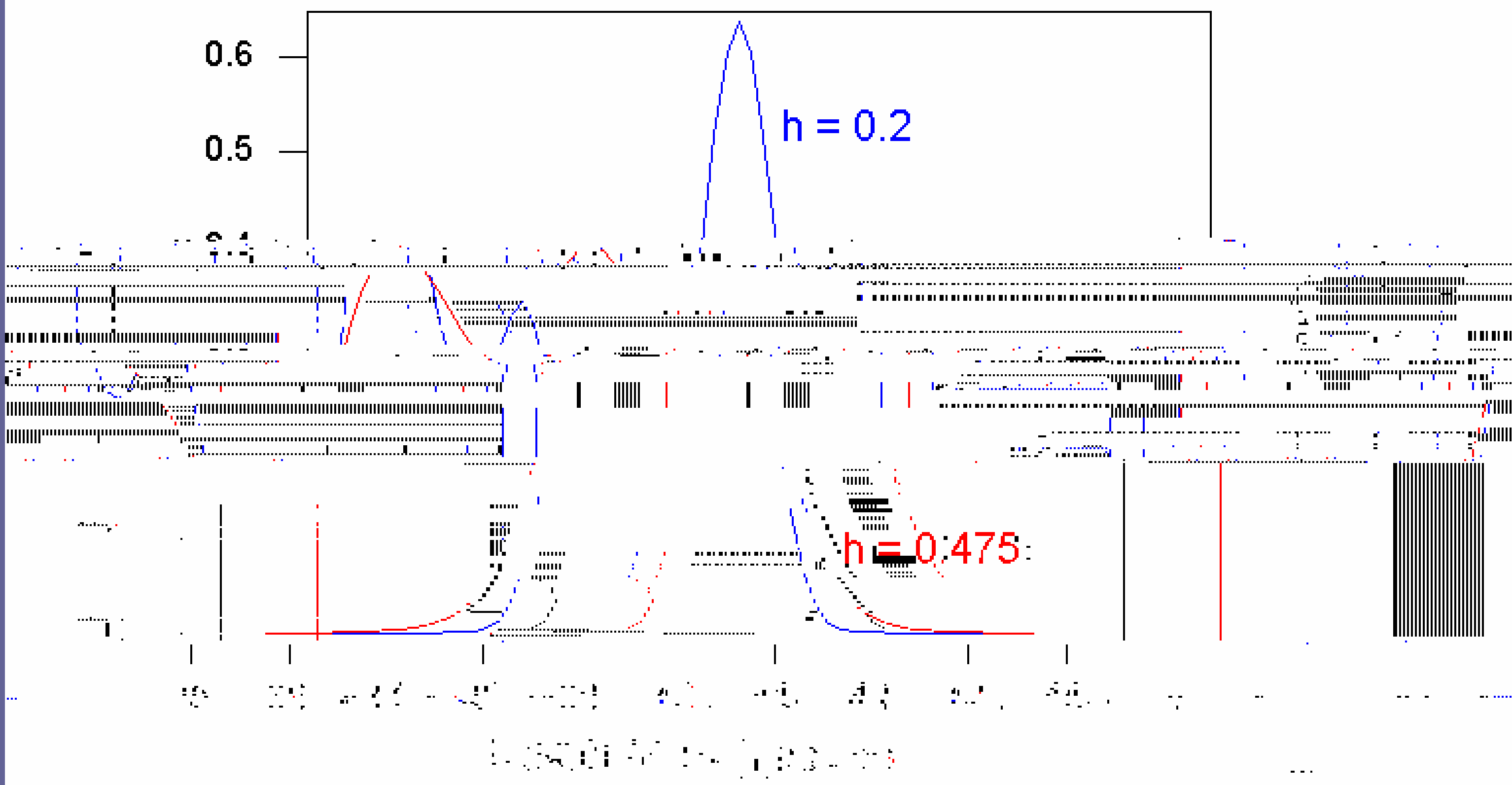


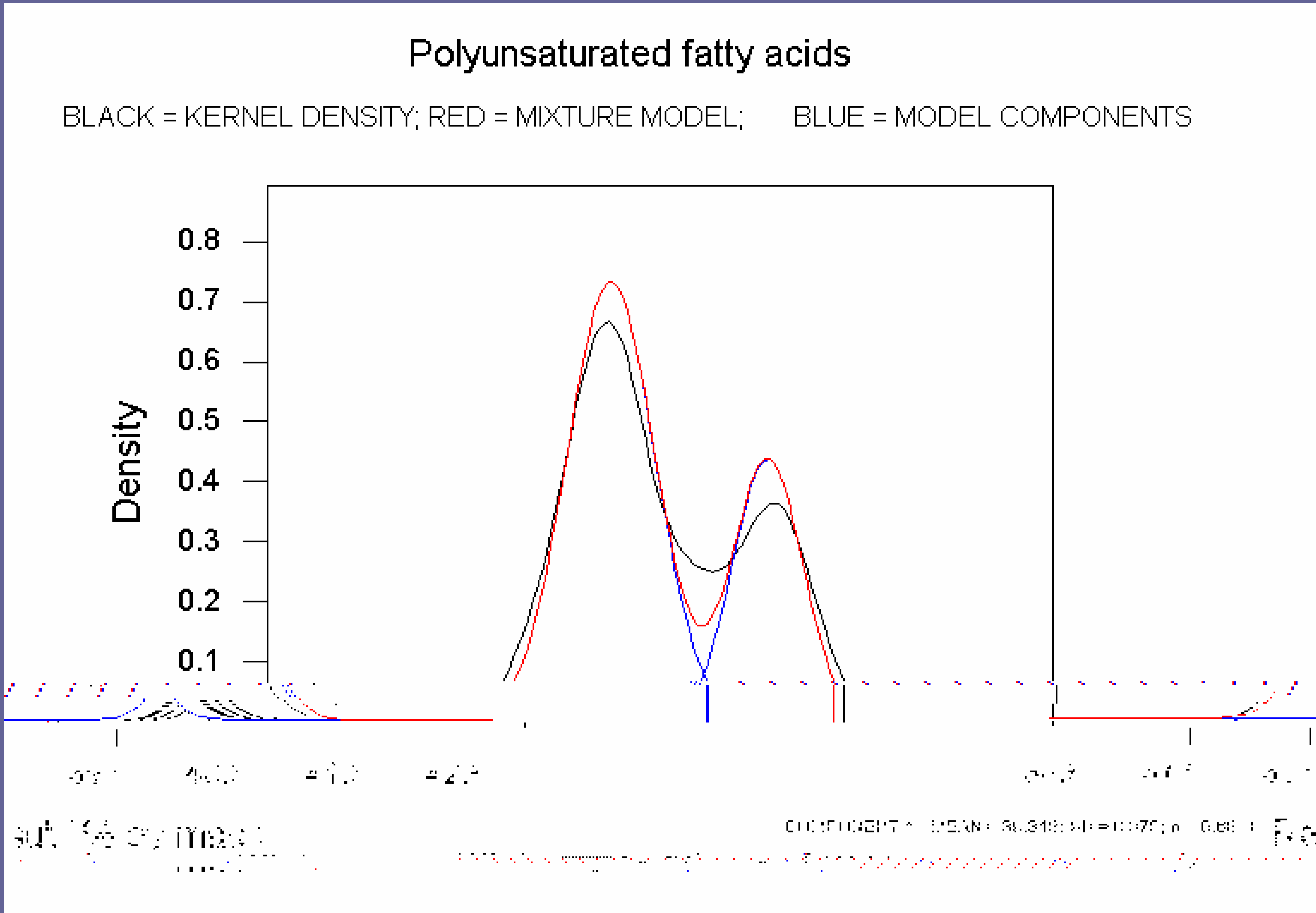
Polyunsaturated fatty acids

Kernel density mode count



Kernel densities--polyunsaturated fatty acids



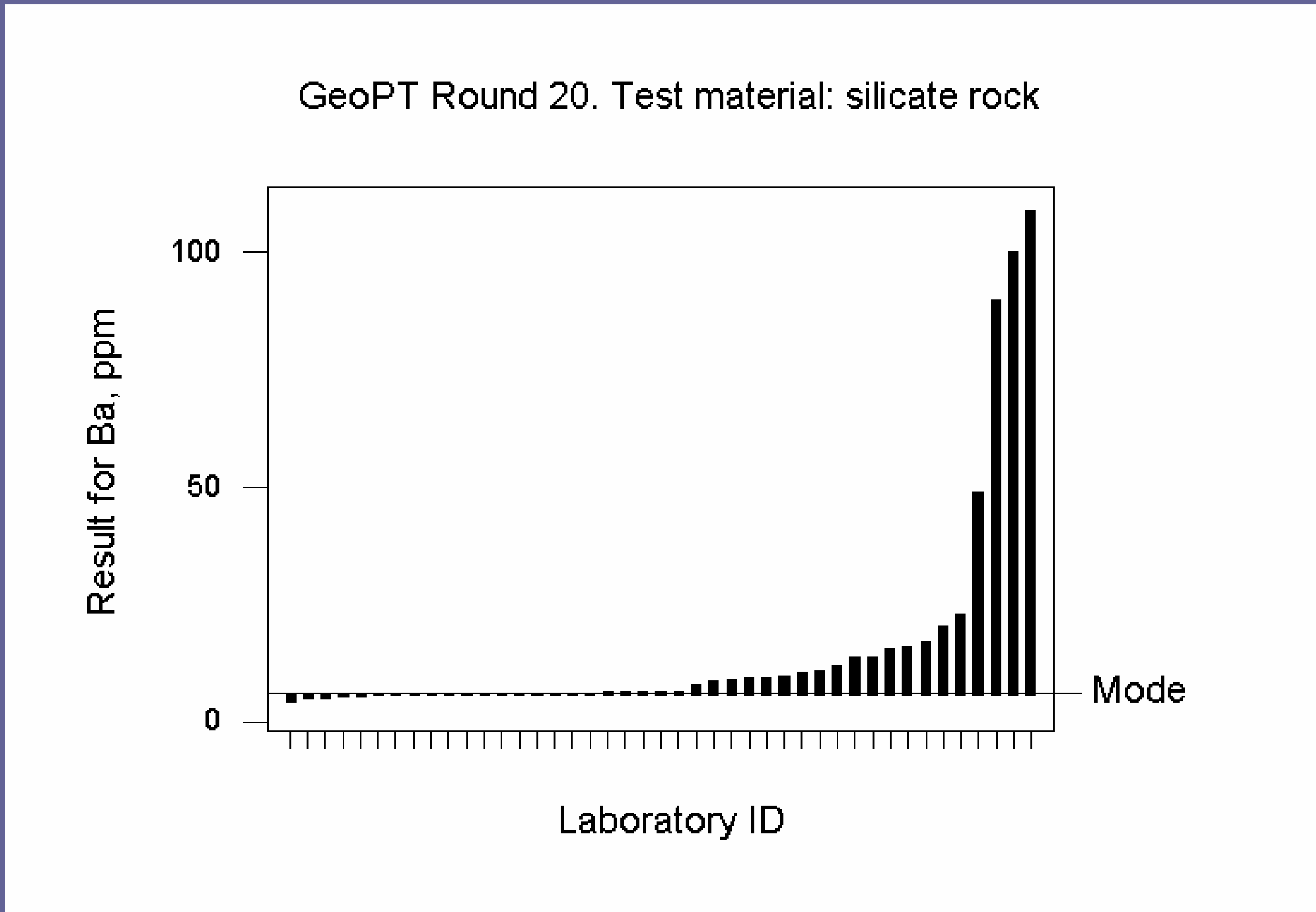


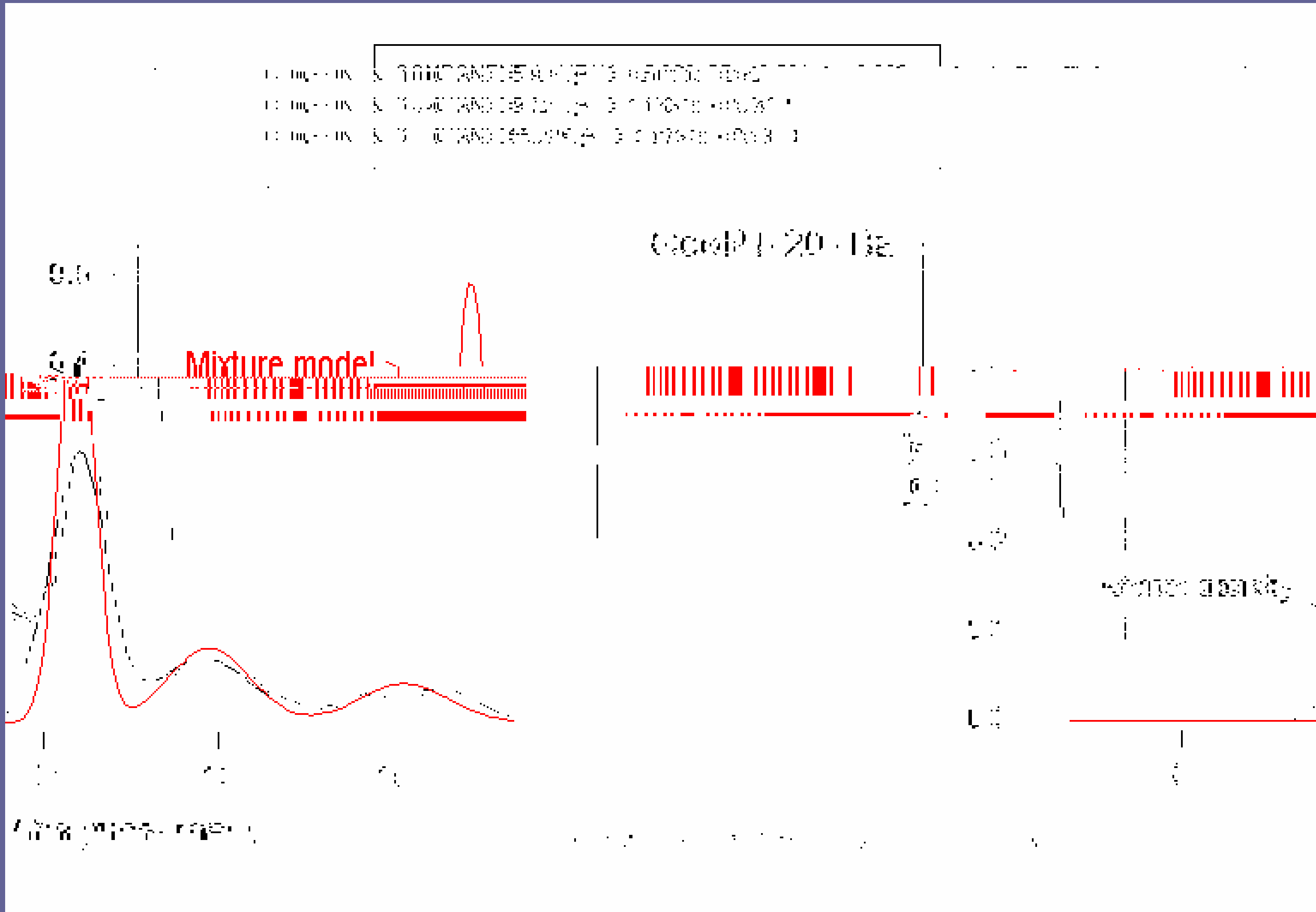
Polyunsaturated fatty acids
BLACK = KERNEL DENSITY; RED = MIXTURE MODEL; BLUE = MODEL COMPONENTS
Density
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

What went wrong?

- Analyte defined as % fatty acid in oil.
- Most labs used an internal standard method.
- Hypothesis: other labs (incorrectly) reported result based on methyl ester peak area ratio.
- Incorrect results expected to be high by a factor of 1.05.
- Ratio of modes found = 1.04.

Example 3—Ba in silicate rock




Mixture model

Self-referential scoring

- Nearly always, more than 90% of laboratories receive a z-score between ±2.
-

What more do we need?

- We need a method that *evaluates* the results in relation to their intended use, rather than merely describing them.
- We need a method in which a score of (say) -3.1 has an meaning independent of the analyte, matrix, or analytical method.
- We need a method based on:

Fitness for purpose

- Fitness for purpose occurs when the uncertainty of the result u_f gives best value for money.
- If the uncertainty is smaller than u_f , the analysis may be too expensive.
- If the uncertainty is larger than u_f , the cost and the probability of a mistaken decision will rise.

Fitness for purpose

- The value of u_f can sometimes be estimated objectively by decision theory methods.
- Usually u_f can be simply agreed between the laboratory and the customer by professional judgement.
- In the proficiency test context, u_f should be

- If we now define a z-score thus:

Conclusions—optimal scoring

- Use z-scores based on fitness for purpose.
- Estimate the consensus as the robust mean and its uncertainty as $\hat{\ }_{rob} \big/ \sqrt{n}$ if the dataset is roughly symmetric.
- If the dataset is skewed and plausibly composite, use a kernel density or a mixture model to find a consensus.

And finally……

- Each dataset is unique. It is impossible to define a sequence of statistical operations that will properly handle every eventuality.

- Statistics (in the right hands) assists, but cannot replace, professional judgement.

General references

- *The International Harmonised Protocol for Proficiency Testing in Analytical Chemistry Laboratories* (revised), M Thompson, S L R Ellison and R Wood. *Pure Appl. Chem*., 2006, **78**, 145-196.

- R E Lawn, M Thompson and R F Walker, *Proficiency testing in analytical chemistry*. The Royal Society of Chemistry, Cambridge, 1997.

- ISO Guide 43. *Proficiency testing by interlaboratory comparisons*, Geneva, 1997.

- ISO Standard 13528. *Statistical methods for use in proficiency testing by interlaboratory comparisons,* Geneva, 2005.