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

$$z = \frac{x - x_A}{\sigma_p}$$

“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 necessarily identical with the true value. PT providers and users have to be alert to this possibility.
- The consensus must have a sufficiently small uncertainty. This usually requires >20 participants.

What exactly *is* a ‘consensus’?

- Mean? - easy to calculate, but affected by outliers and asymmetry.
- Robust mean? - fairly easy to calculate, handles outliers but affected by strong asymmetry.
- Median? - easy to calculate, more robust for asymmetric distributions, but larger standard error than robust mean.
- Mode? - intuitively good, handles strong skews, difficult to define, difficult to calculate.

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

$$\hat{\mu}_{rob},\ \hat{\sigma}_{rob}$$

- 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.

Huber's H15 estimators

$$\mathbf{x}^{\mathrm{T}} = \left[x_1 \quad x_2 \quad \Lambda \quad x_n\right]$$

$$\text{Set} \quad 1 < k < 2, \quad p = 0, \;\; \hat{\mu}_0 = \text{median}, \;\; \hat{\sigma}_0 = 1.5 \times \text{MAD}$$

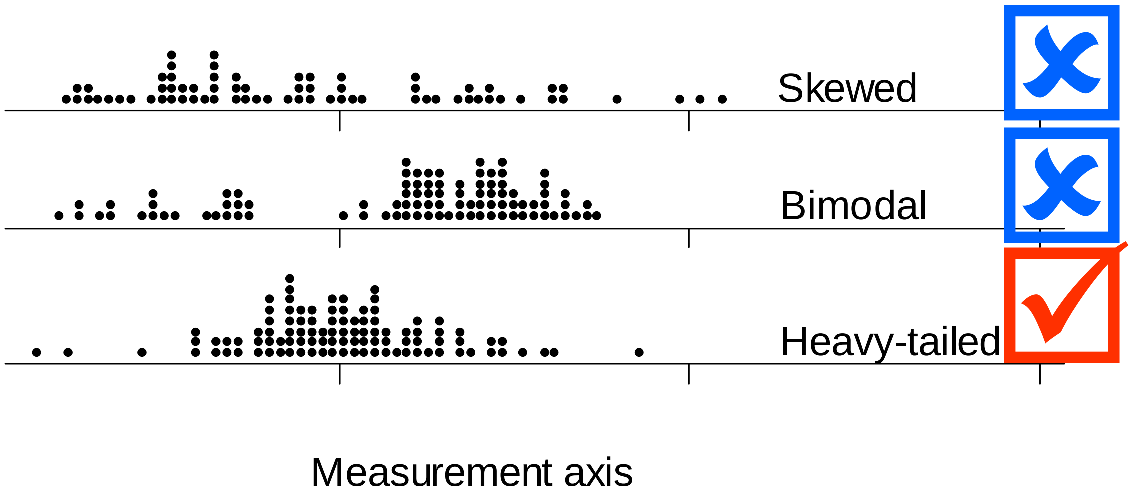
$$\tilde{x}_i = \begin{cases} x_i & \text{if} \quad \hat{\mu}_p - k\hat{\sigma}_p < x_i < \hat{\mu}_p + k\hat{\sigma}_p \\ \hat{\mu}_p - k\hat{\sigma}_p & \text{if} \quad x_i < \hat{\mu}_p - k\hat{\sigma}_p \\ \hat{\mu}_p + k\hat{\sigma}_p & \text{if} \quad x_i < \hat{\mu}_p + k\hat{\sigma}_p \end{cases}$$

$$\hat{\mu}_{p+1} = \text{mean}(\tilde{x}_i)$$

$$\hat{\sigma}^2_{p+1} = f(k)\,\text{var}(\tilde{x}_i)$$

$$\text{If not converged,} \quad p = p + 1$$

When can I safely use robust estimates?



The robust mean as consensus

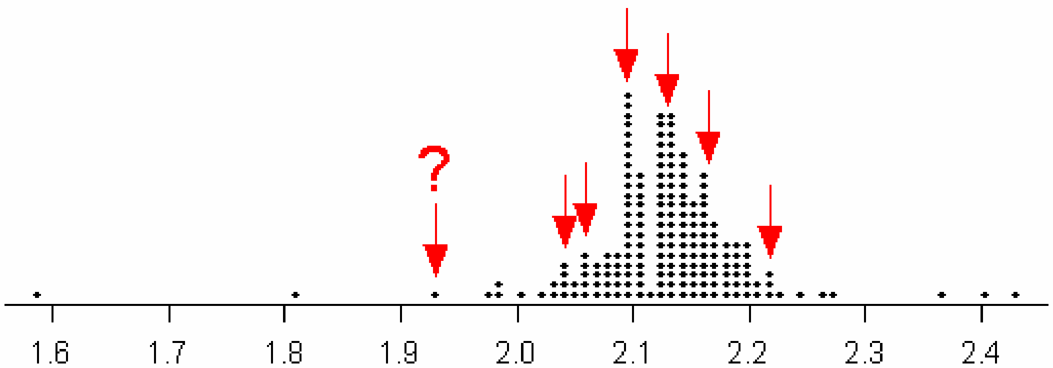
- The robust mean provides a useful consensus in the great majority of instances.
- The uncertainty of this consensus can be safely taken as $u(x_a) = \hat{\sigma}_{rob} / \sqrt{n}$

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?



The normal kernel density for identifying a mode

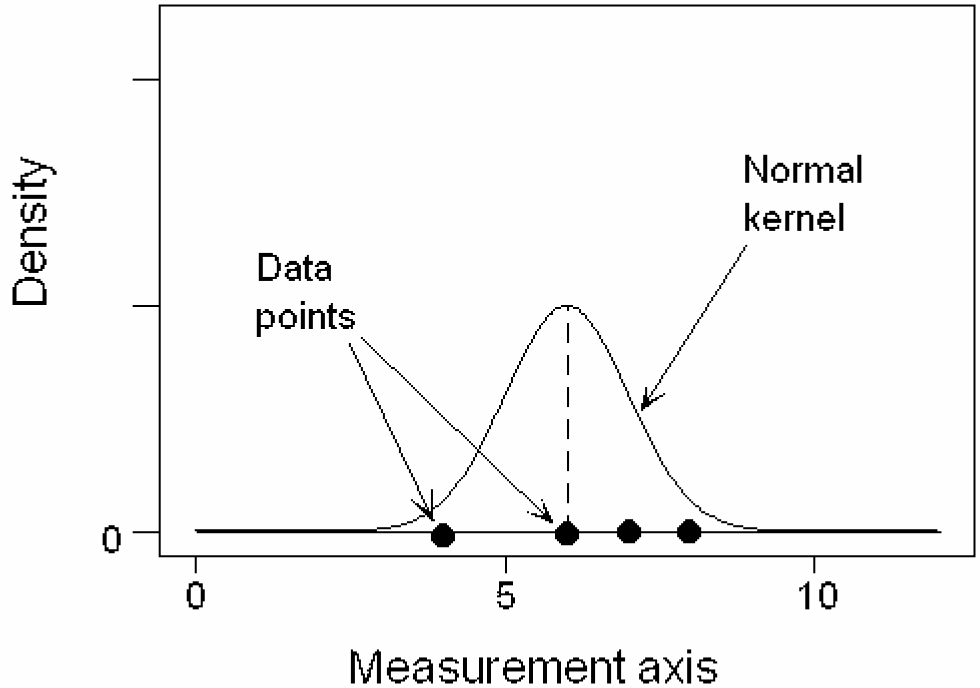
$$y = \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{x - x_i}{h}\right)$$

where *Φ* is the standard normal density,

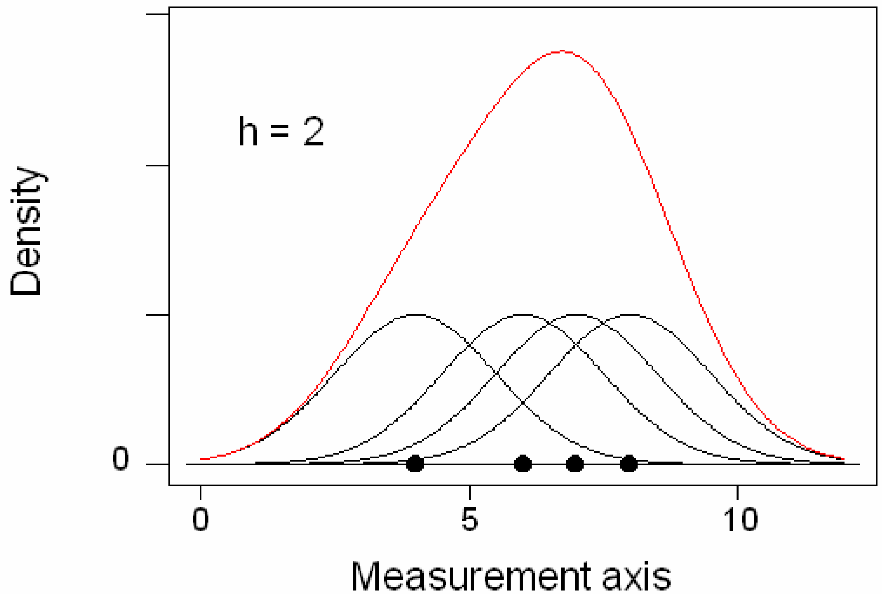
$$\phi(a) = \frac{\exp(-a^2/2)}{\sqrt{2\pi}}$$

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

A normal kernel

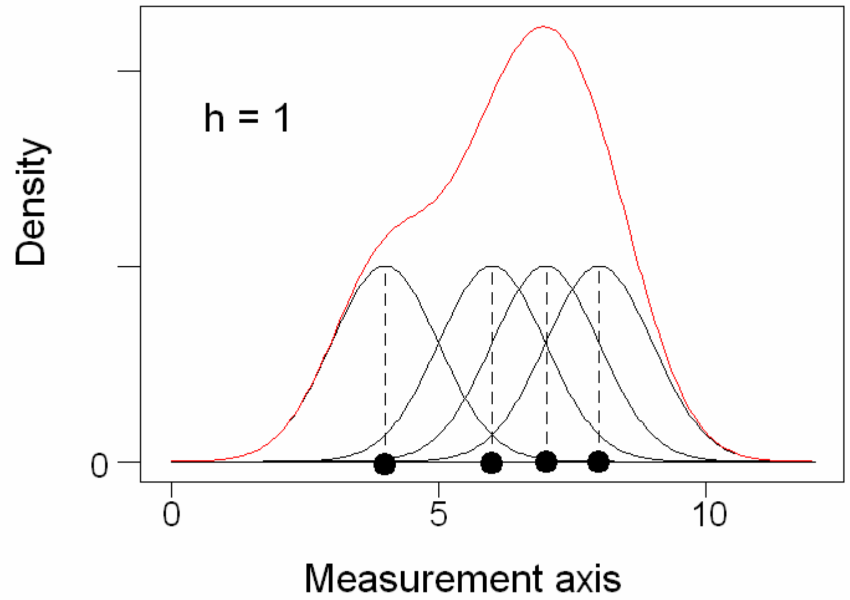


A kernel density



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.

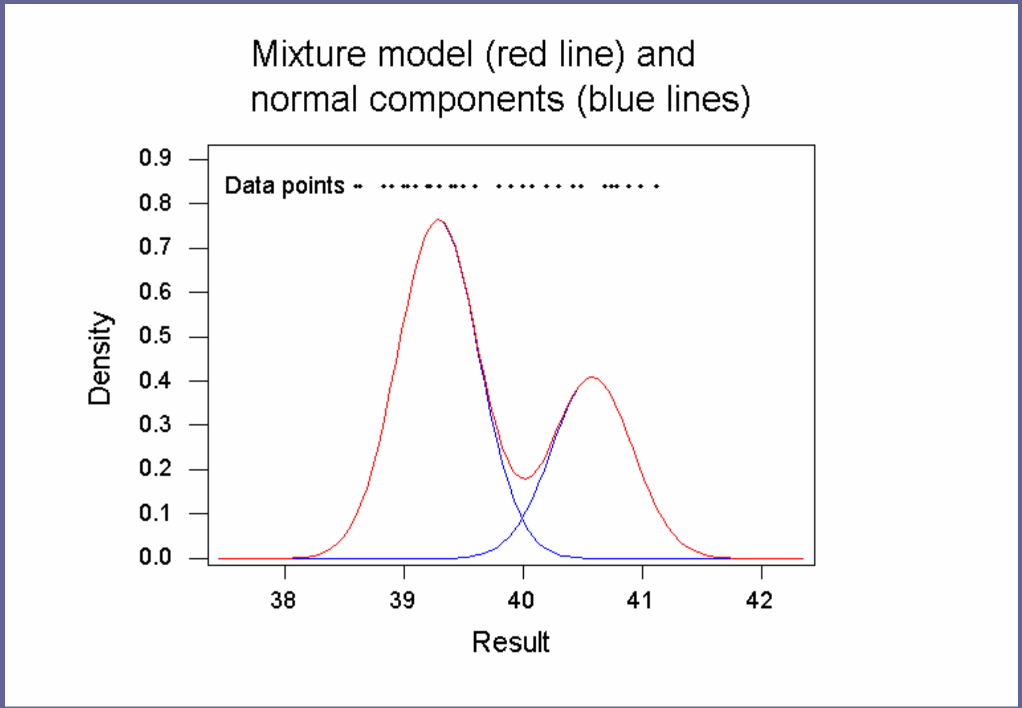
- The bootstrap is a general procedure, based on resampling, for estimating standard errors of complex statistics.

- 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

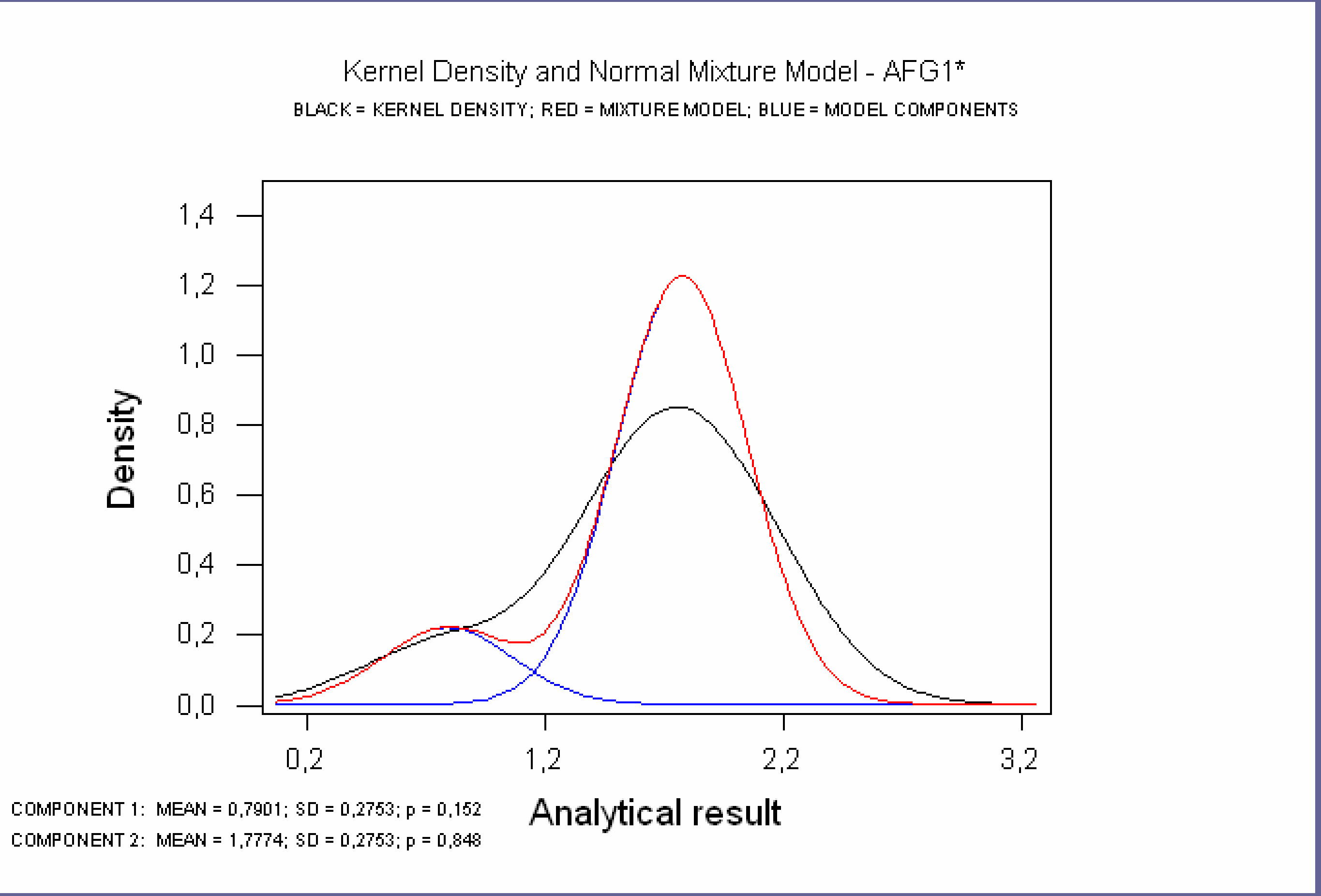
Mixture models and consensus



- For each component you can calculate:

  - a mean
  - a variance
  - a proportion

2-component normal mixture model and kernel density



The normal mixture model

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

$$f_j(y) = \frac{\exp(-(y-\mu_j)^2 / 2\sigma^2}{\sqrt{2\pi}\sigma}$$

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

Mixture models found by the maximum likelihood method (the EM algorithm)

- The M-step

$$\hat{p}_j = \sum_{i=1}^{n} \hat{P}(j \mid y_i) / n$$

$$\hat{\mu}_j = \sum_{i=1}^{n} y_i \hat{P}(j \mid y_i) \Big/ \sum_{i=1}^{n} \hat{P}(j \mid y_i)$$

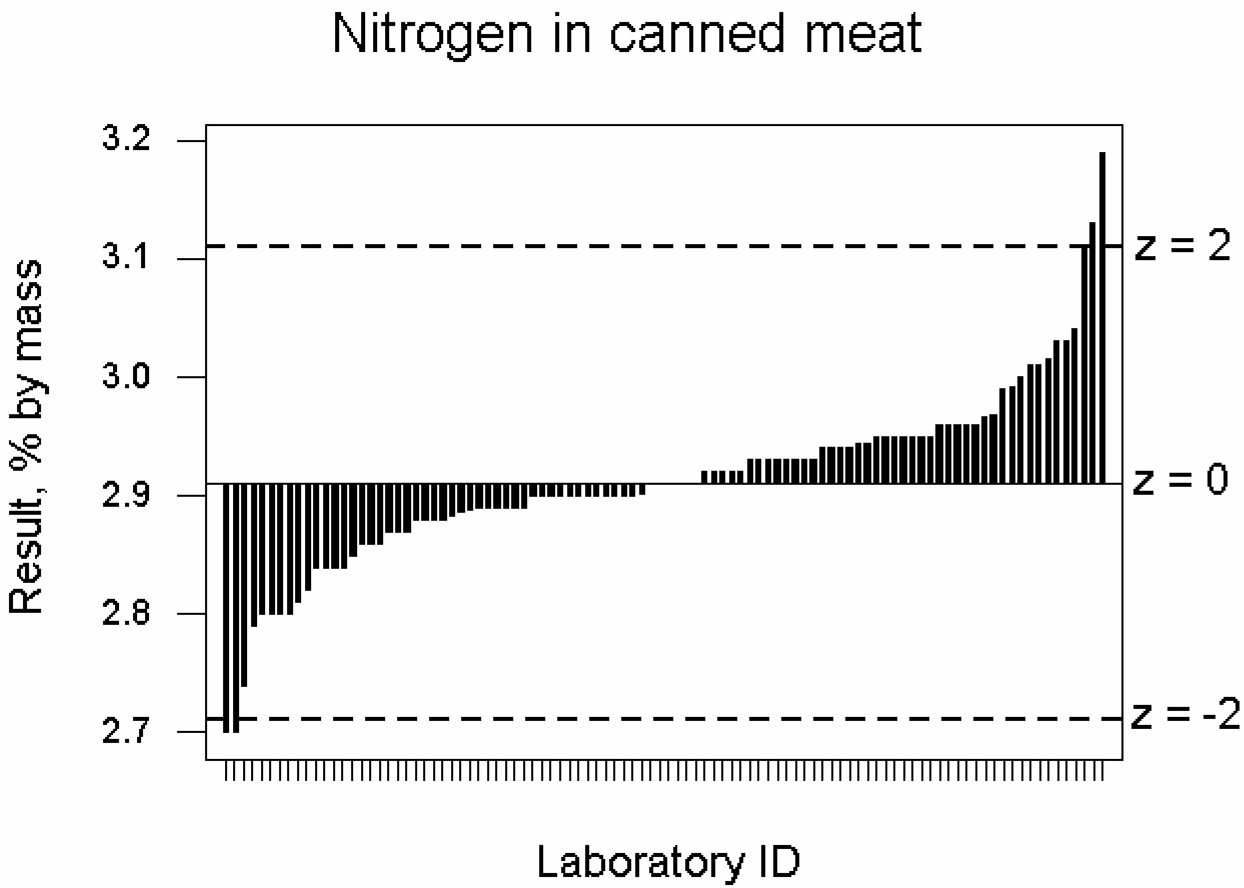
$$\hat{\sigma}^2 = \sum_{j=1}^{m} \sum_{i=1}^{n} \left((y_i - \hat{\mu}_j)^2 \hat{P}(j \mid y_i) \right) \Big/ \hat{P}(j \mid y_i)$$

- The E-step

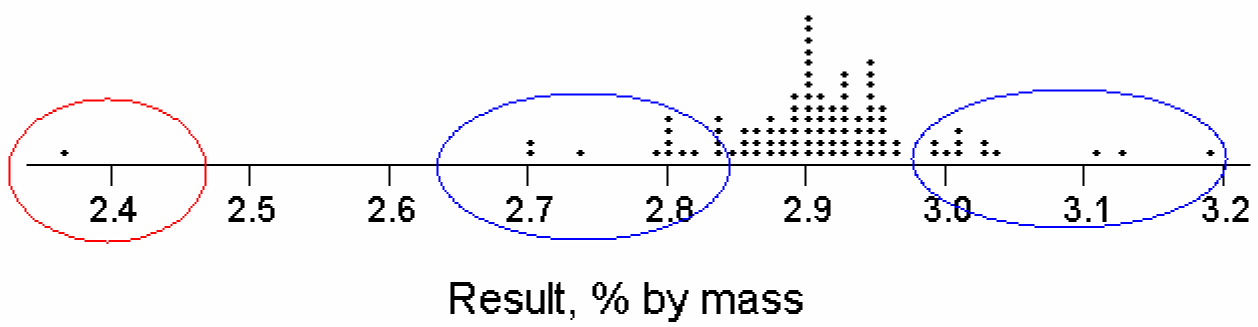
$$\hat{P}(j \mid y_i) = \hat{p}_j f_j(y_i) \Big/ \sum_{j=1}^{m} \hat{p}_j f_j(y_i)$$

Example datasets

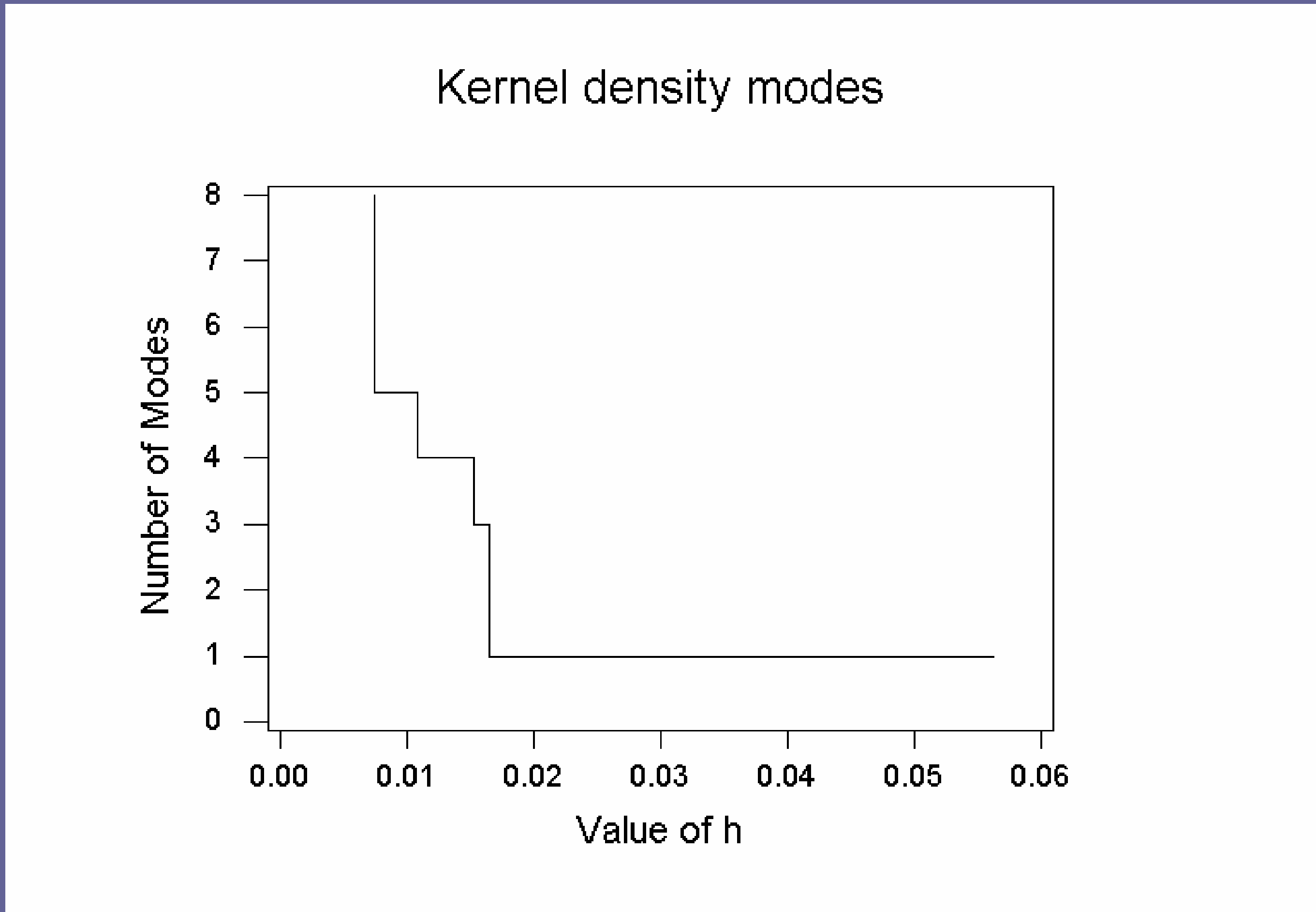
Example dataset 1

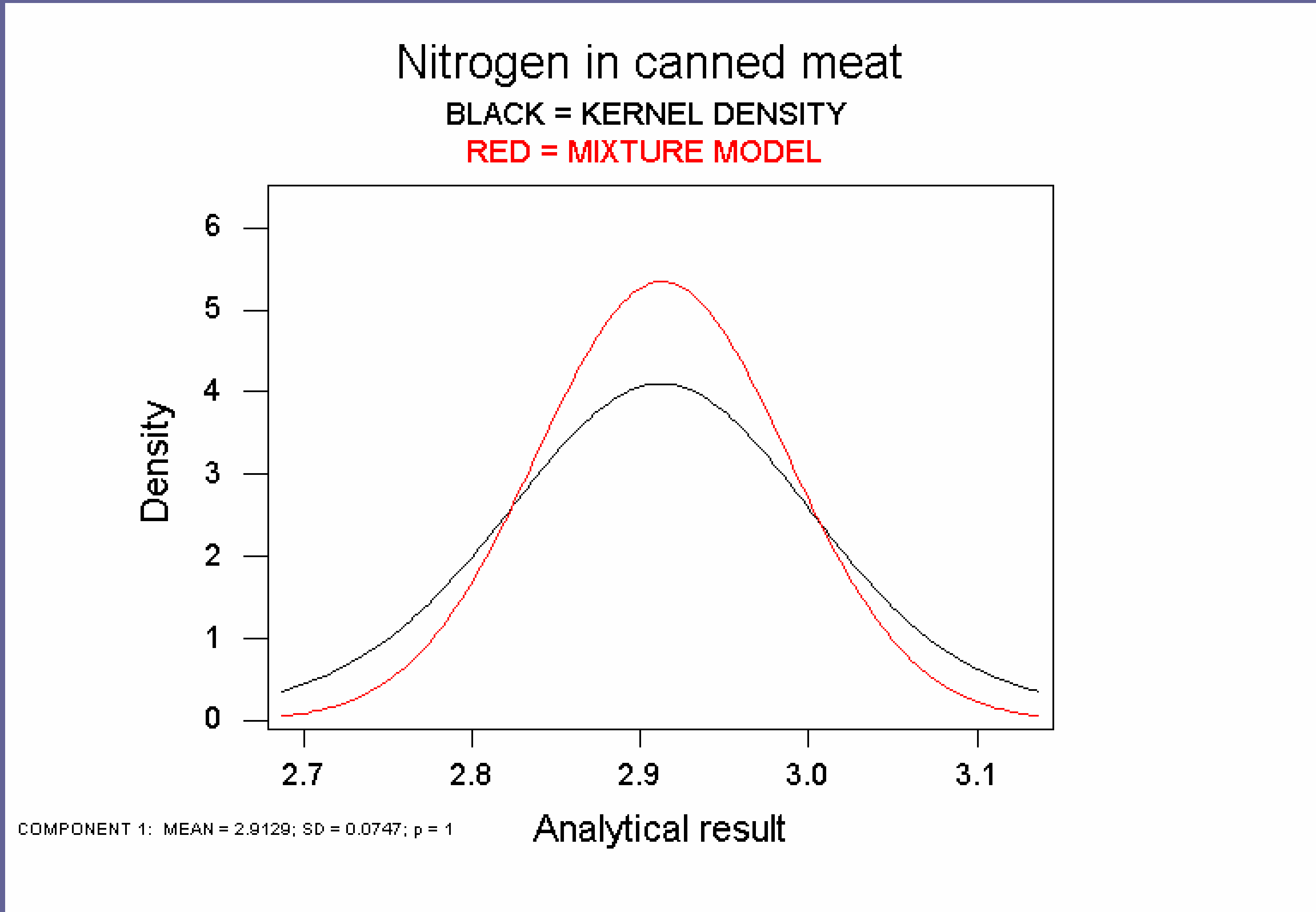


Nitrogen in canned meat



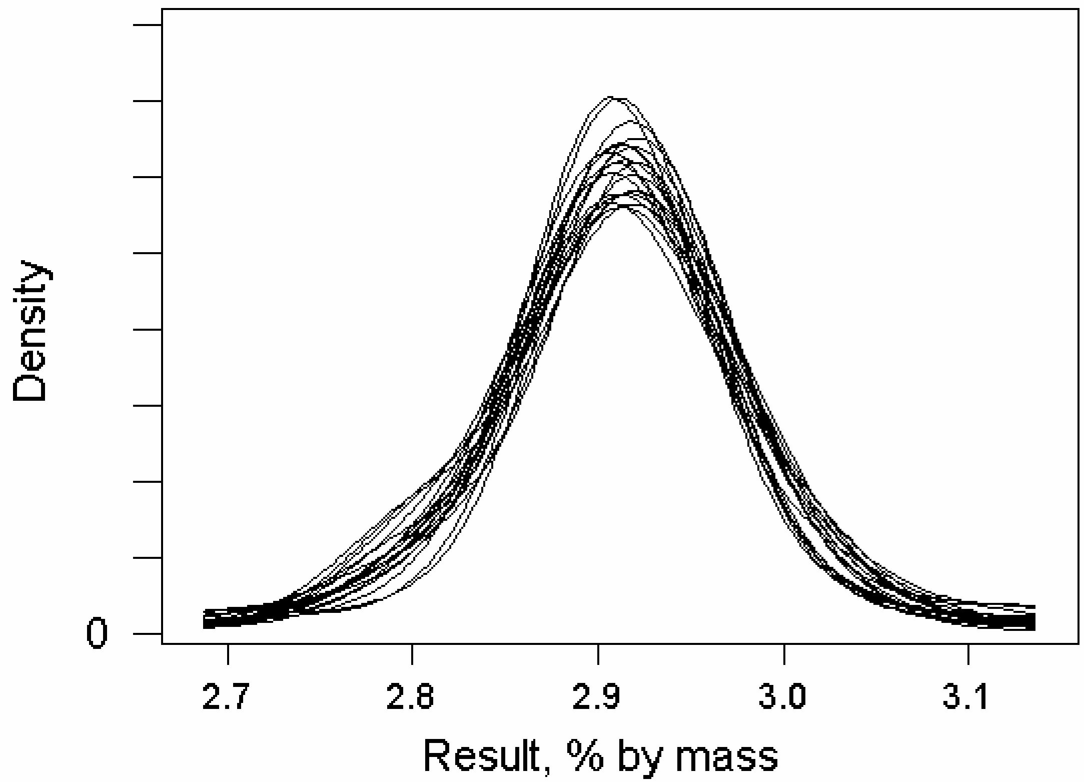
Number of modes vs smoothing factor *h*





Nitrogen in canned meat
BLACK = KERNEL DENSITY
RED = MIXTURE MODEL
Density
0
1
2
3
4
5
6
2.7
2.8
2.9
3.0
3.1
Analytical result
COMPONENT 1: MEAN = 2.9129; SD = 0.0747; p = 1

Bootstrapped kernel density plots



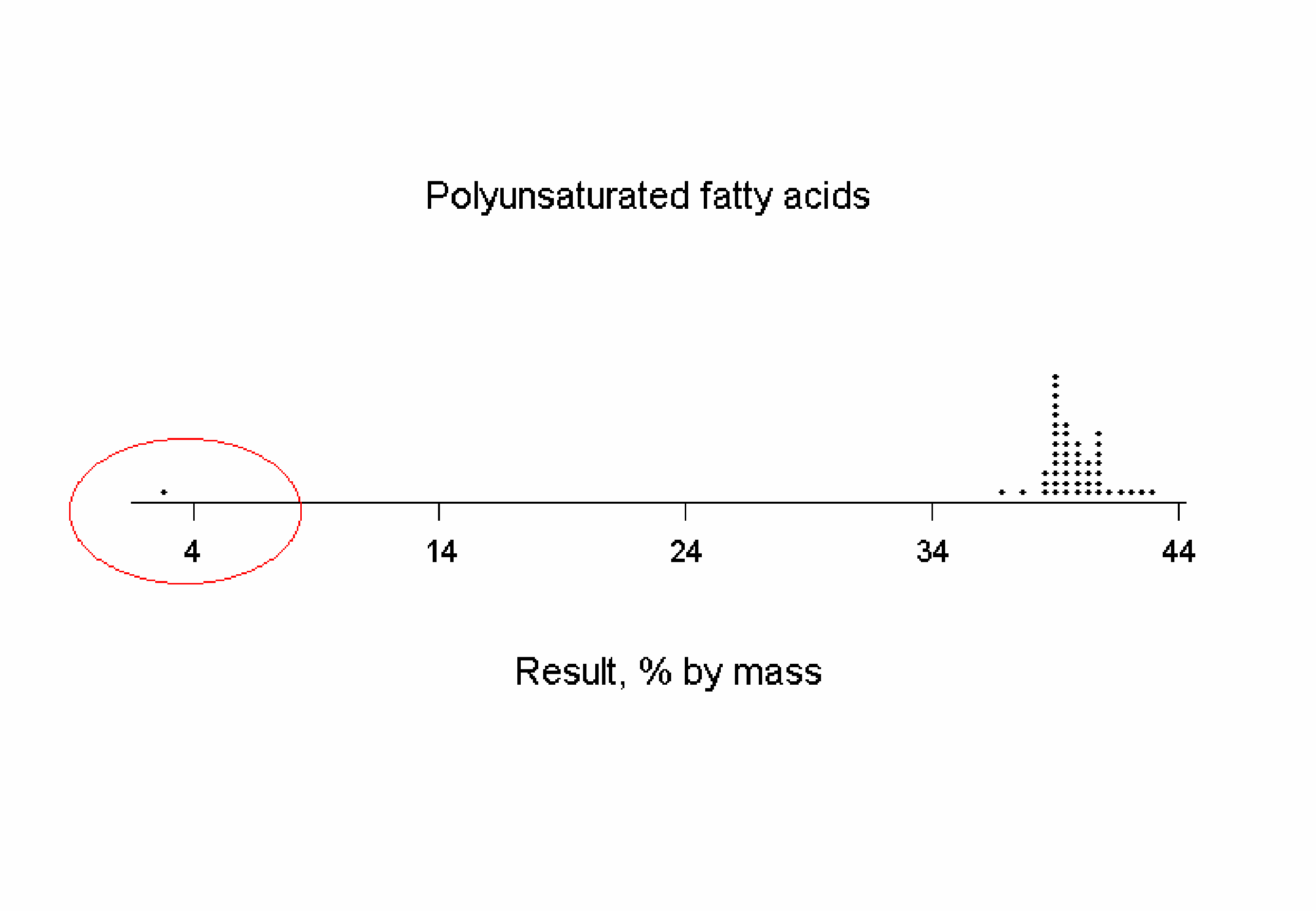
Statistics: dataset 1

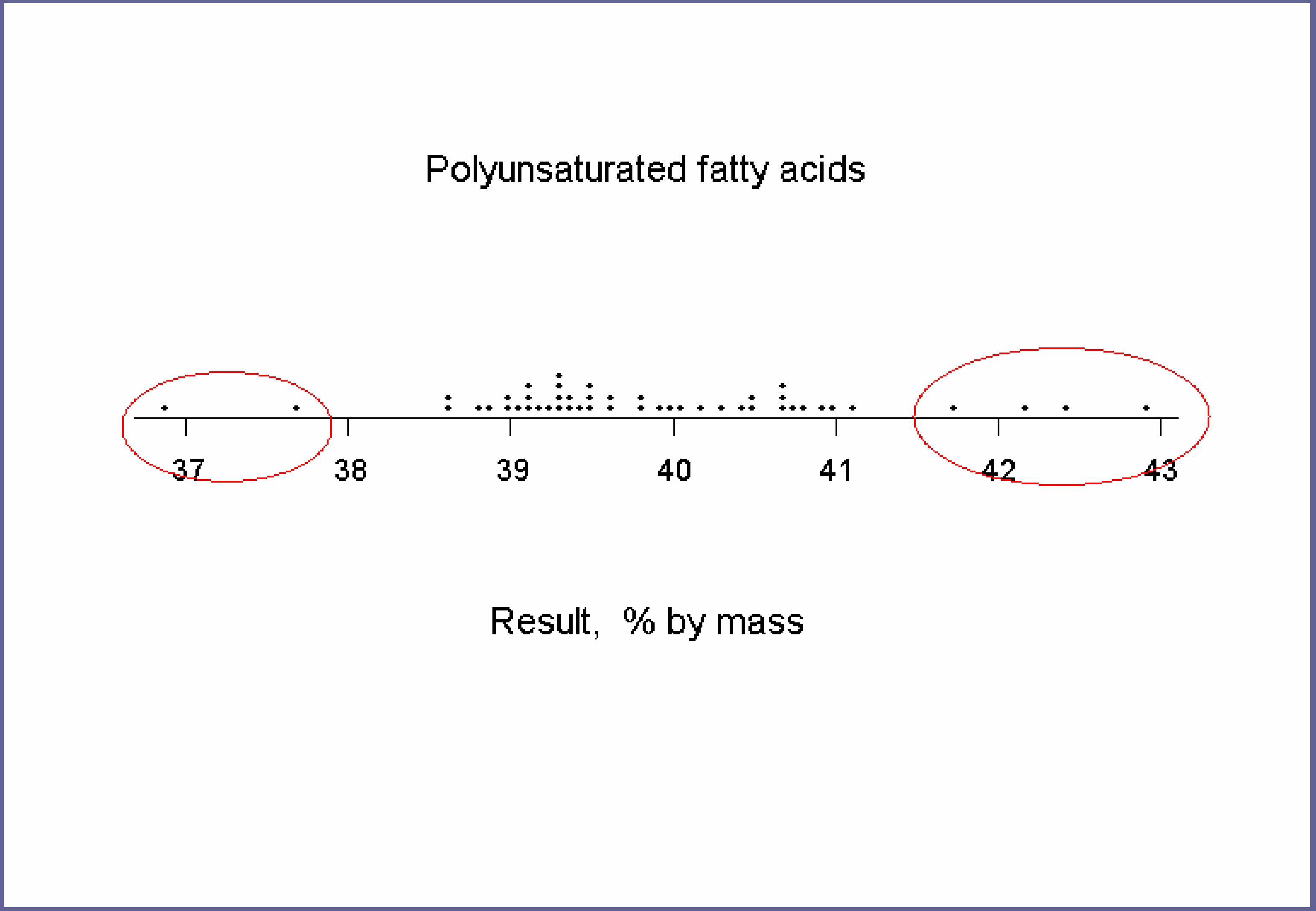
	$\hat{\mu}$	$\hat{\sigma}$	$se(\hat{\mu})$
Robust	2.912	0.056	0.0056
Kernel density mode	2.912	-	0.0056
Mixture model	2.913	0.075	0.0075

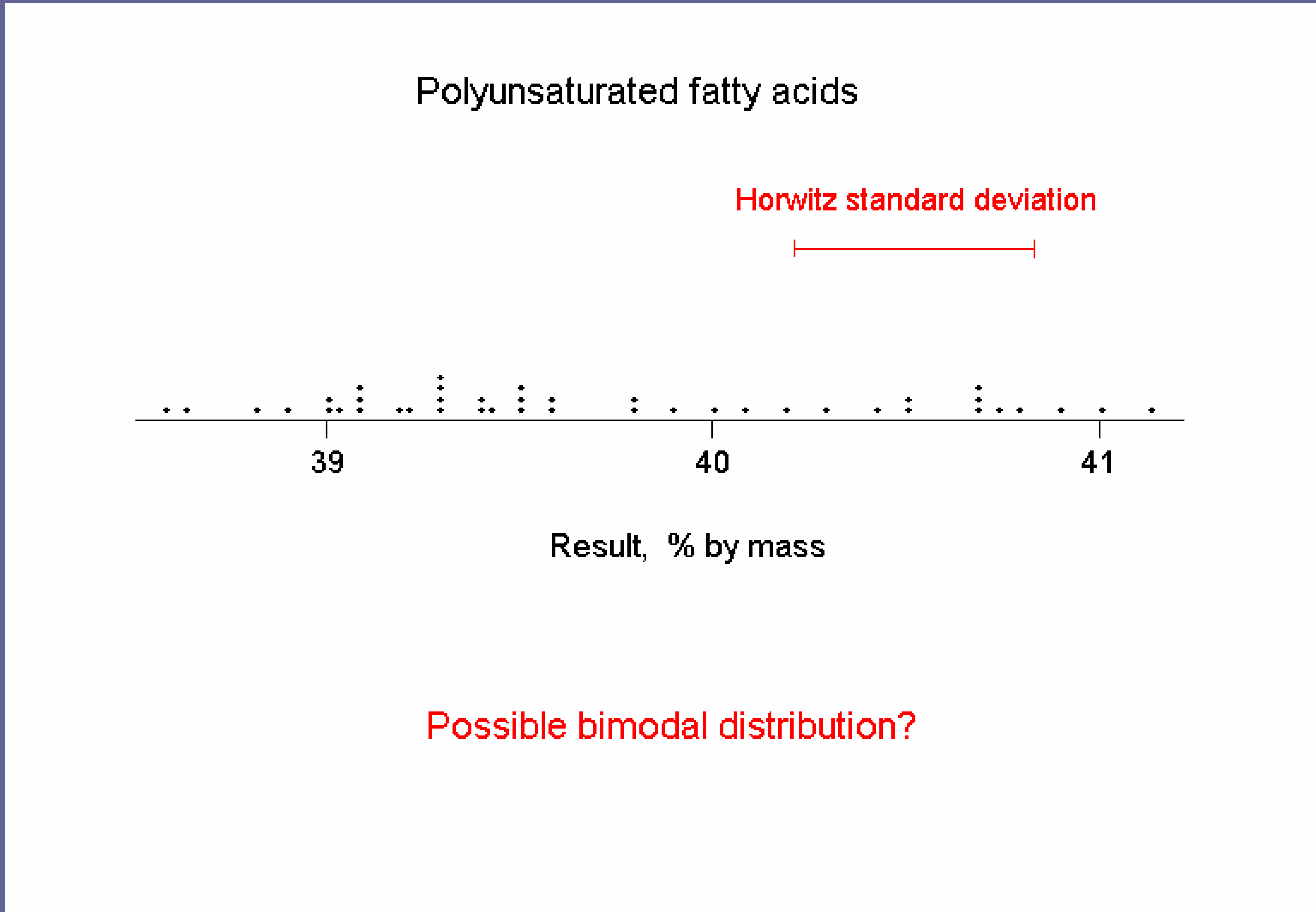
Skewed/multimodal distributions

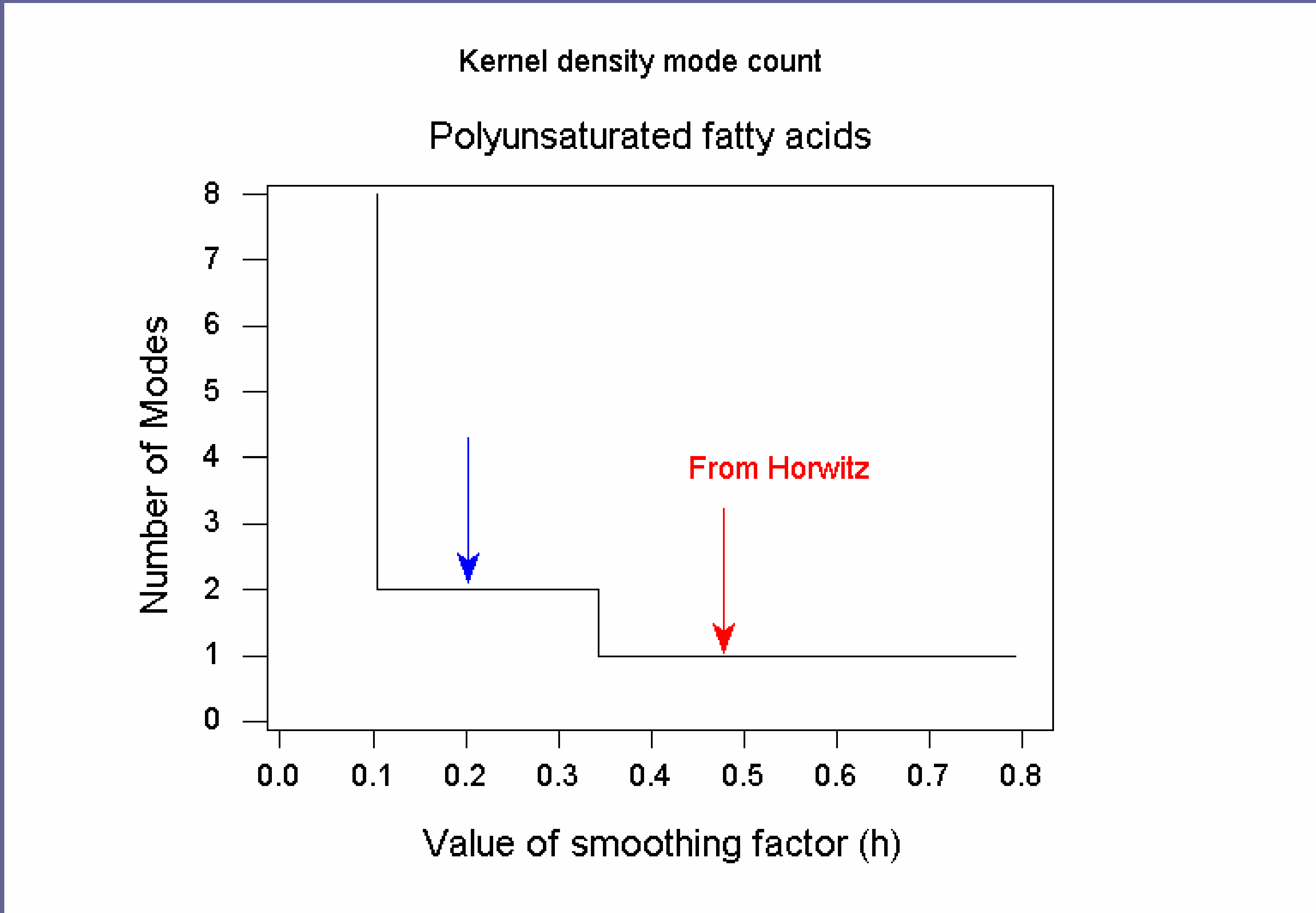
- Skews and extra modes can arise when the participants' results come from two or more inconsistent methods.
- Skews can also arise as an artefact at low concentrations of analyte as a result of common data recording practices.
- Rarely, skews can arise when the distribution is truly lognormal (*e.g.*,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

Example dataset 2

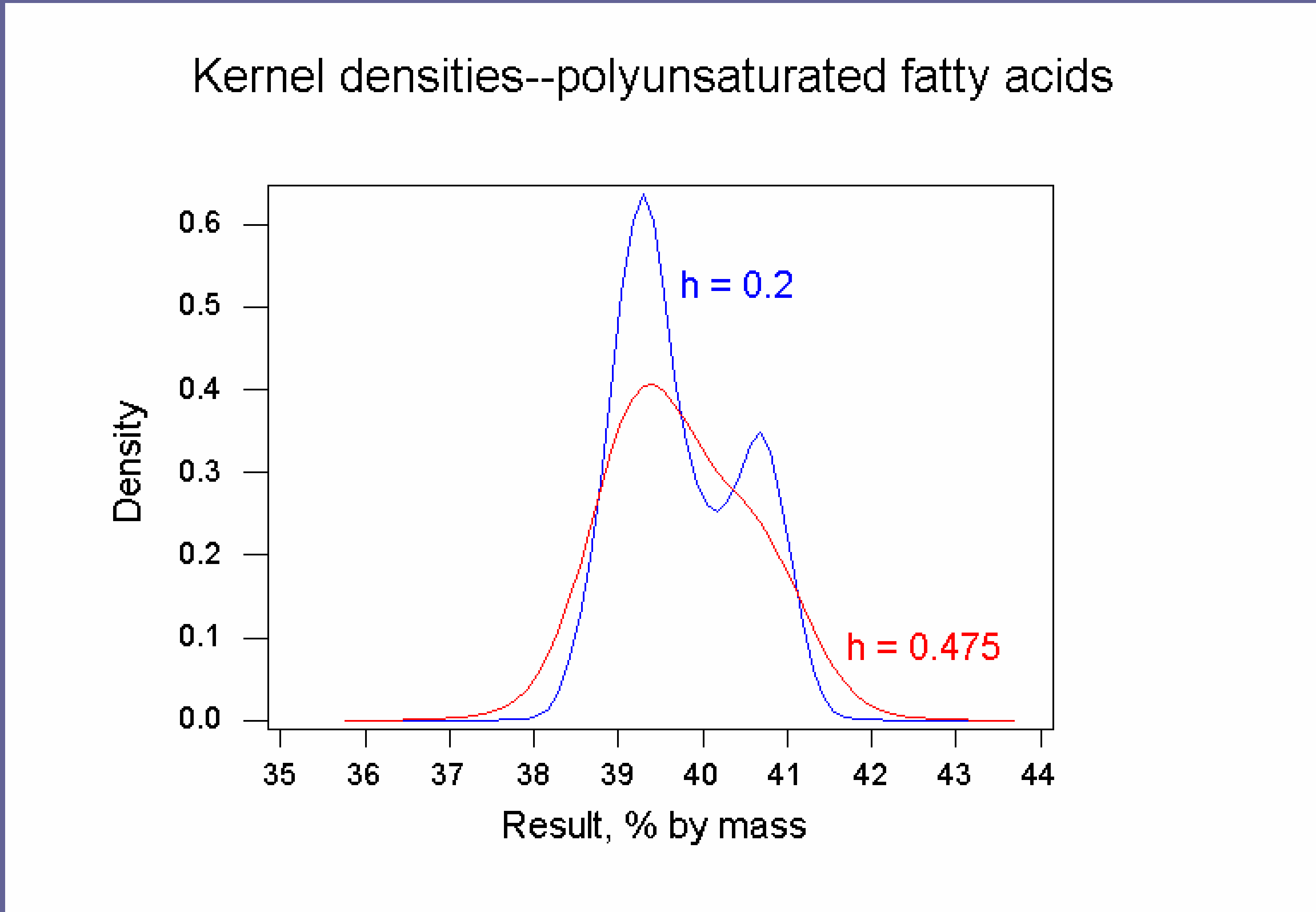


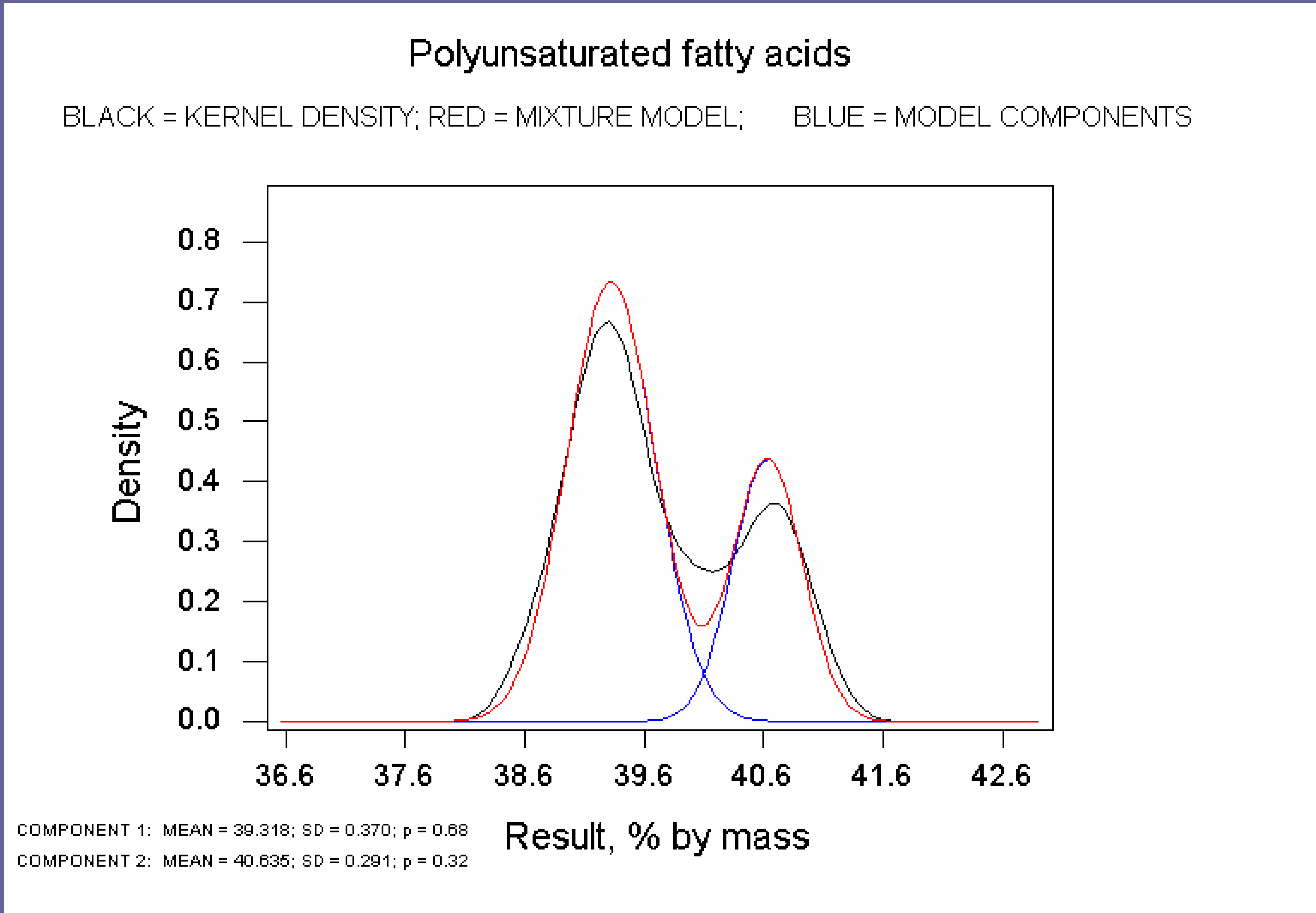

Polyunsaturated fatty acids
37
38
39
40
41
42
43
Result, % by mass


Polyunsaturated fatty acids
Horwitz standard deviation
39
40
41
Result, % by mass
Possible bimodal distribution?



Kernel density mode count
Polyunsaturated fatty acids
Number of Modes
0
1
2
3
4
5
6
7
8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Value of smoothing factor (h)
From Horwitz


Kernel densities--polyunsaturated fatty acids
h = 0.2
h = 0.475
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
35
36
37
38
39
40
41
42
43
44
Result, % by mass

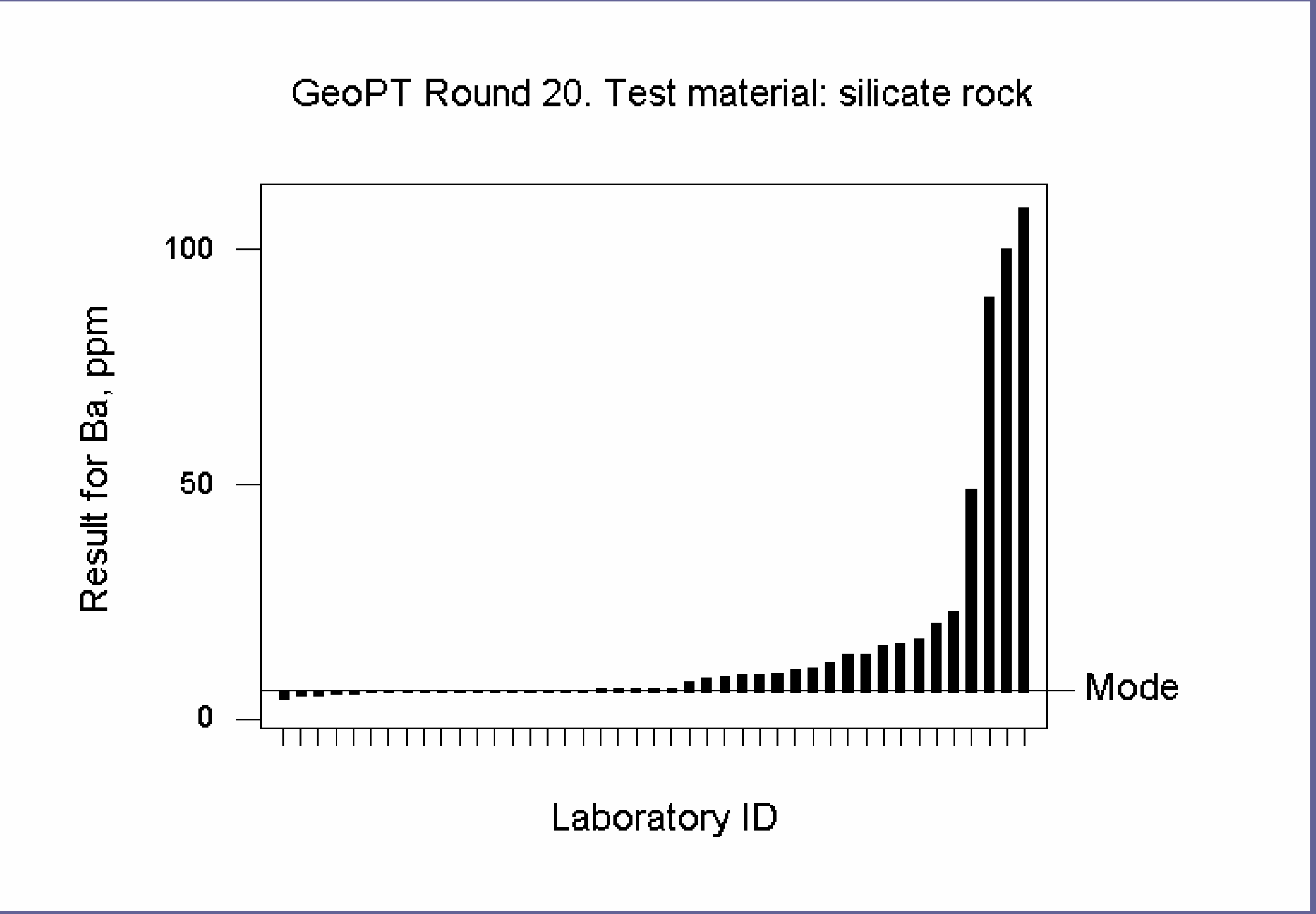


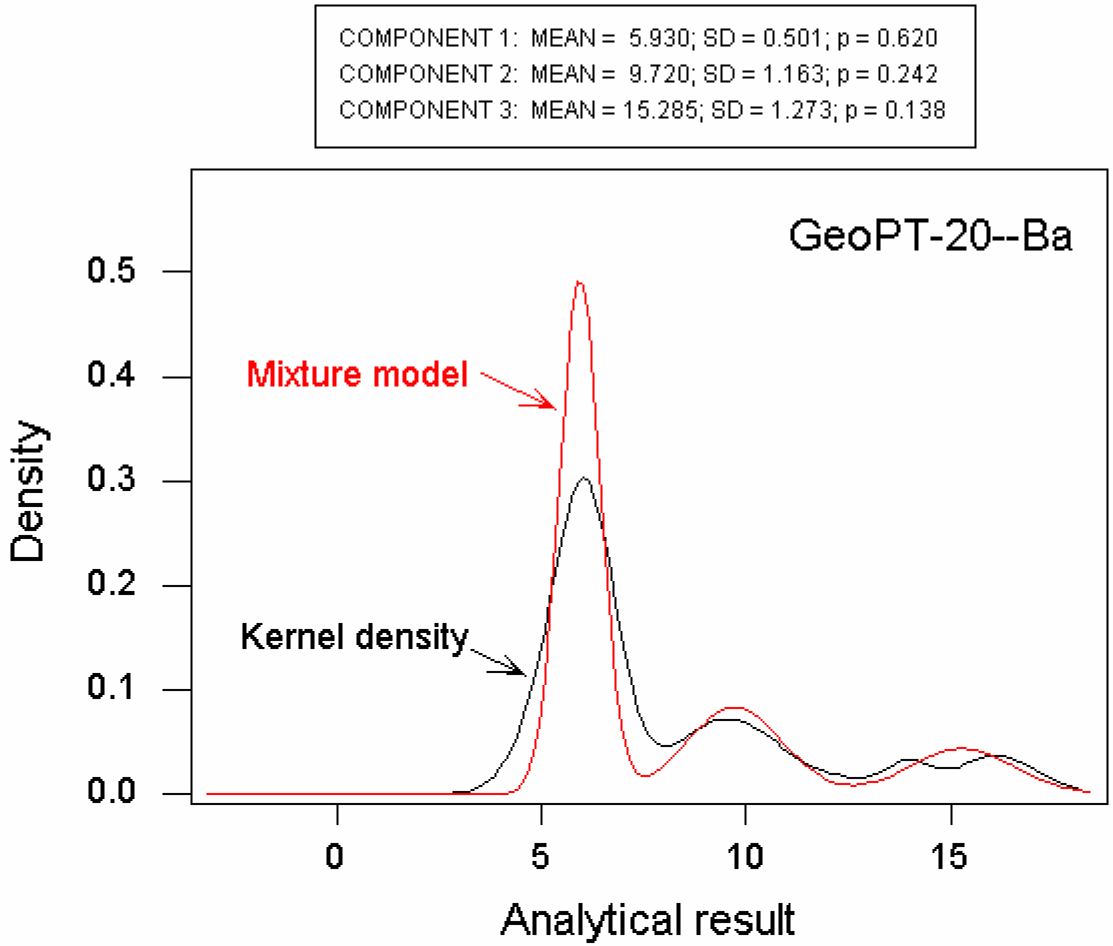
Polyunsaturated fatty acids
BLACK = KERNEL DENSITY; RED = MIXTURE MODEL; BLUE = MODEL COMPONENTS
Density
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
36.6
37.6
38.6
39.6
40.6
41.6
42.6
Result, % by mass
COMPONENT 1: MEAN = 39.318; SD = 0.370; p = 0.68
COMPONENT 2: MEAN = 40.635; SD = 0.291; p = 0.32

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





COMPONENT 1: MEAN = 5.930; SD = 0.501; p = 0.620
COMPONENT 2: MEAN = 9.720; SD = 1.163; p = 0.242
COMPONENT 3: MEAN = 15.285; SD = 1.273; p = 0.138
GeoPT-20--Ba
Mixture model
Kernel density
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
Analytical result

Choice of value for σ_p

- Robust standard deviation of participants' results in round?
- From perception of how well similar methods perform?
- Legislation?
- Other?

Self-referential scoring

$$z = \left(x - \hat{\mu}_{rob}\right) / \hat{\sigma}_{rob}$$

- Nearly always, more than 90% of laboratories receive a z-score between ±2.
- This suggests, to both provider and participants, that accuracy is generally OK, whether or not that is the case.
- No reference is made to end-user requirements.
- z-Scores for a participant cannot be meaningfully compared round-to-round.

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

- 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 determined by the scheme provider.

Reference: T Fearn, S A Fisher, M Thompson, and S L R Ellison, *Analyst*, 2002, **127**, 818-824.

A score that meets all of the criteria

- If we now define a z-score thus:

  $$z = \left(x - \hat{\mu}_{rob}\right) / \sigma_p \quad \text{where} \quad \sigma_p \equiv u_f$$

  we have a z-score that is both robustified against extreme values *and* tells us about fitness for purpose.
- In an exactly compliant laboratory, scores of 2<|z|<3 will be encountered occasionally, and scores of |z|>3 rarely.
- Better performers will receive fewer of these extreme z-scores, worse performers more.

Conclusions—optimal scoring

- Use z-scores based on fitness for purpose.
- Estimate the consensus as the robust mean and its uncertainty as $\hat{\sigma}_{rob} / \sqrt{n}$ 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.

Statistical References

- **Mixture models**
  M Thompson. *Accred Qual Assur*. 2006, **10**, 501-505.
  AMC Technical Brief No. 23, 2006.   www/rsc.org/amc
- **Kernel densities**
  B W Silverman, *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.
  AMC Technical Brief, no. 4, 2001   www/rsc.org/amc
- **The bootstrap**
  B Efron and R J Tibshirani, *An introduction to the bootstrap.* Chapman and Hall, London, 1993
  AMC Technical Brief, No. 8, 2001   www/rsc.org/amc
- **Robust statistics**
  Analytical Methods Committee, *Analyst*,1989, **114**, 1489
  AMC Technical Brief No 6, 2001 (www/rsc.org/amc)
  P J Rousseeuw, *J. Chemomet*, 1991, **5**, 1.

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.