Statistics and Data Analysis in Proficiency Testing

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

Where do we use statistics in proficiency testing?

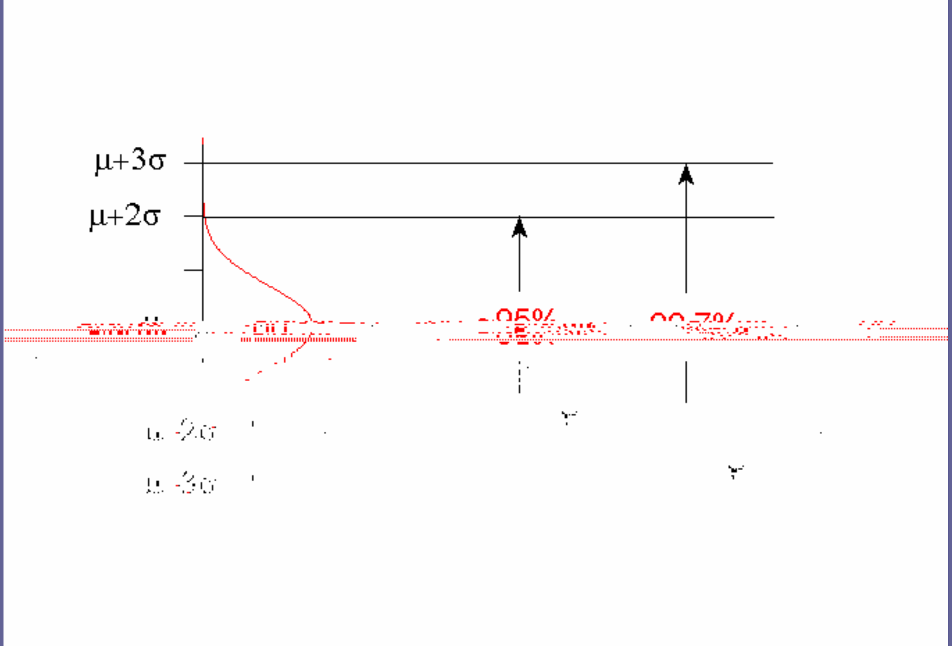
- Finding a consensus and its uncertainty to use as an assigned value
- Assessing participants' results
- Assessing the efficacy of the PT scheme
- Testing for sufficient homogeneity and stability of the distributed test material
- Others

Criteria for an ideal scoring method

- Adds value to raw results.
- Easily understandable, based on the properties of the normal distribution.
- Has no arbitrary scaling transformation.
- Is transferable between different concentrations, analytes, matrices, and measurement principles.

How can we construct a score?

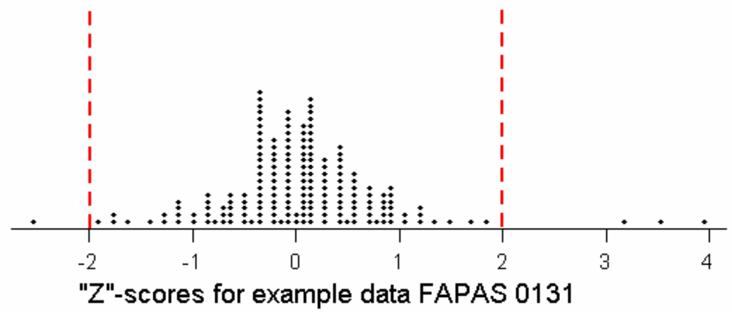
- An obvious idea is to utilise the properties of the normal distribution to interpret the results of a proficiency test.



BUT…

We do not make any assumptions about the actual data.

A weak scoring method



97% of scores in range $-2 < z < 2$

$$z \quad x \quad \bar{x} \big/ s$$

$$\bar{x} \quad 2.126$$

$$s \quad 0.077$$

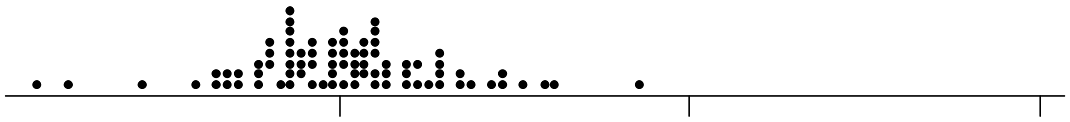
- On average, slightly more than 95% of laboratories receive z-score within the range ±2.

Robust mean and standard deviation

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

- 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.

Can I use robust estimates?



Huber's H15

$\mathbf{x}^{\mathrm{T}} \quad x_1 \quad x_2 \qquad x_n$

Set $1 \quad k \quad 2, \quad p \quad 0, \quad \hat{\ }_0 \quad$ median, $\quad \hat{\ }_0 \quad 1.5 \quad$ MAD

$$\tilde{x}_i \quad \begin{matrix} x_i & \text{if} & \hat{\ }_p \;\; k \; \hat{\ }_p \;\; x_i \;\; \hat{\ }_p \;\; k \; \hat{\ }_p \\ \hat{\ }_p \;\; k \; \hat{\ }_p & \text{if} & x_i \;\; \hat{\ }_p \;\; k \; \hat{\ }_p \\ \hat{\ }_p \;\; k \; \hat{\ }_p & \text{if} & x_i \;\; \hat{\ }_p \;\; k \; \hat{\ }_p \end{matrix}$$

$$\hat{\ }_{p\ 1} \quad \text{mean}(\tilde{x}_i)$$

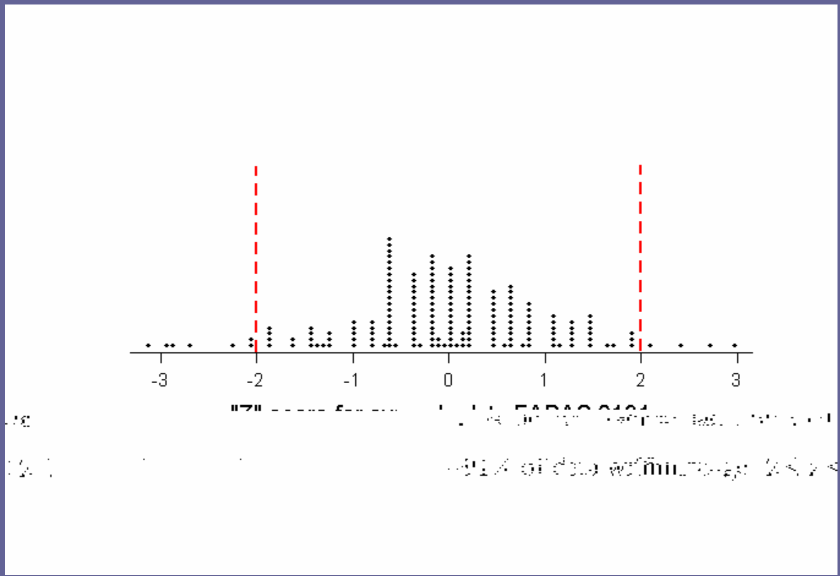
$$\hat{\ }^2_{p\ 1} \quad f(k)\,\text{var}(\tilde{x}_i)$$

If not converged, $\quad p \quad p \quad 1$

References: robust statistics

- Analytical Methods Committee, *Analyst*,1989, **114**, 1489
- AMC Technical Brief No 6, 2001 (download from www/rsc.org/amc)
- P J Rousseeuw, *J. Chemomet*, 1991, **5**, 1.

Is that enough?



$$z \quad x \quad {}_{rob} / \hat{}_{rob}$$

$$\hat{}_{rob} \quad 2.128$$

$$\hat{}_{rob} \quad 0.048$$

- On average, slightly less than 95% of laboratories receive a z-score between ±2.

What more do we need?

- We need a method that *evaluates* the data in relation to its intended use, rather than merely describing it.
- This adds value to the data rather than simply summarising it.
- The method is 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 theoretic methods, but is most often 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:

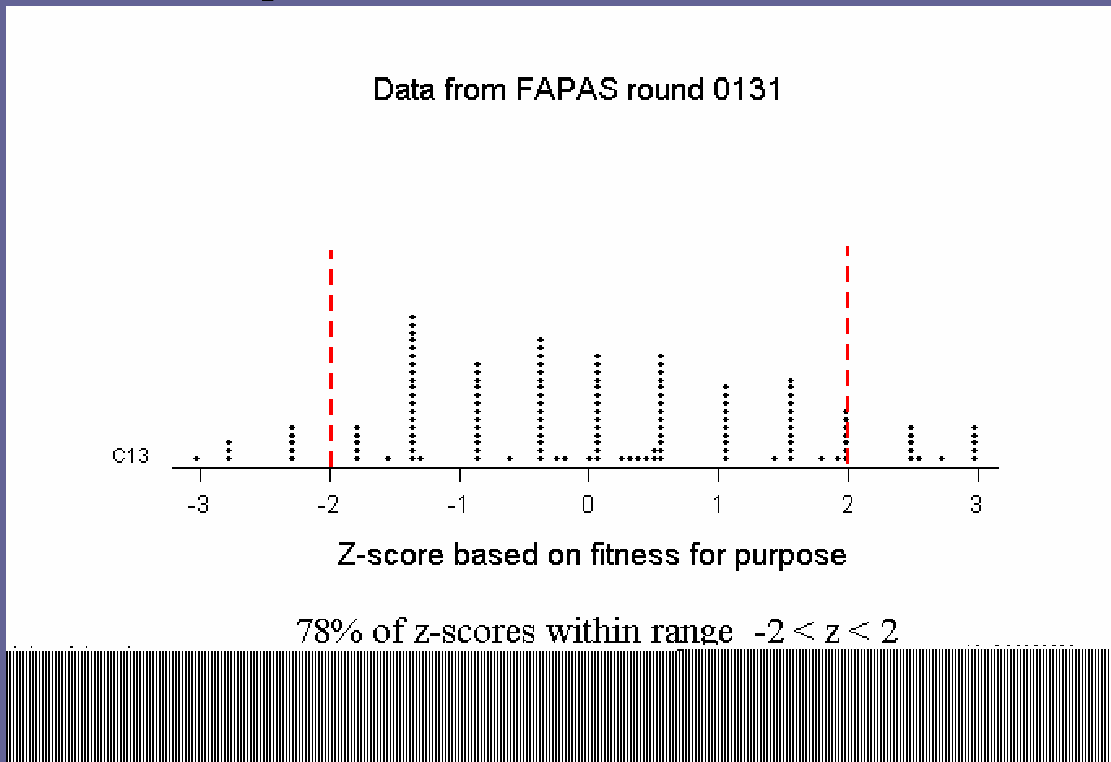
$$z \quad x \quad {}_{rob} / {}_{p} \quad \text{where} \quad {}_{p} \quad u_f$$

  we have a z-score that is both robustified against extreme values *and* tells us something 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.

Example data A again

- Suppose that the fitness for purpose criterion set for the analysis is an RSD of 1%. This gives us:

$$_p \quad 0.01 \quad 2.1 \quad 0.021$$



Finding a consensus from participants' results

- The consensus is not theoretically the best option for the assigned value but is usually the only practicable value.
- The consensus is not necessarily identical with the true value. PT providers have to be alert to this possibility.

What is a ‘consensus’?

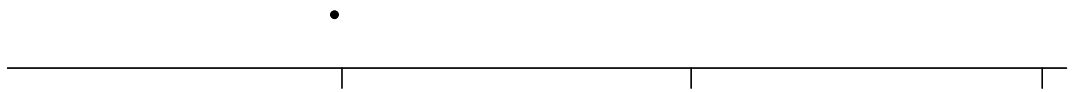
- Mean? - easy to calculate, but affected by outliers and asymmetry.
- Robust mean? - fairly easy to calculate, handles outliers but affected by asymmetry.
- Median? - easy to calculate, more robust for asymmetric distributions, but larger standard error than robust mean.
- Mode? - intuitively good,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

The robust mean as consensus

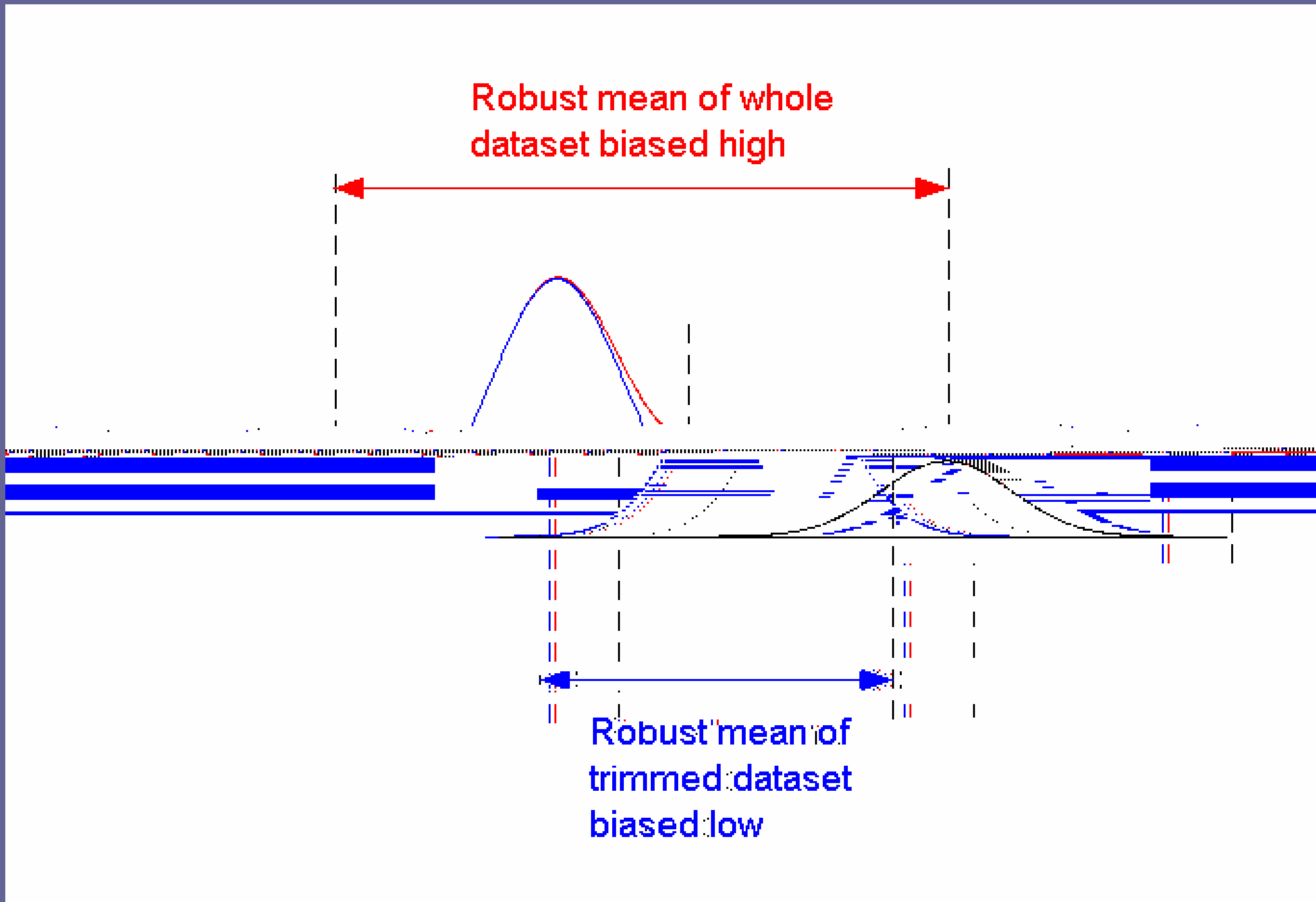
- The robust mean provides a useful consensus in the great majority of instances, where the underlying distribution is roughly symmetric and there are 0-10% outliers.
- The uncertainty of this consensus can be safely taken as

$$u \quad x_a \qquad \hat{\ }_{rob} / \sqrt{n}$$

When can I use robust estimates?

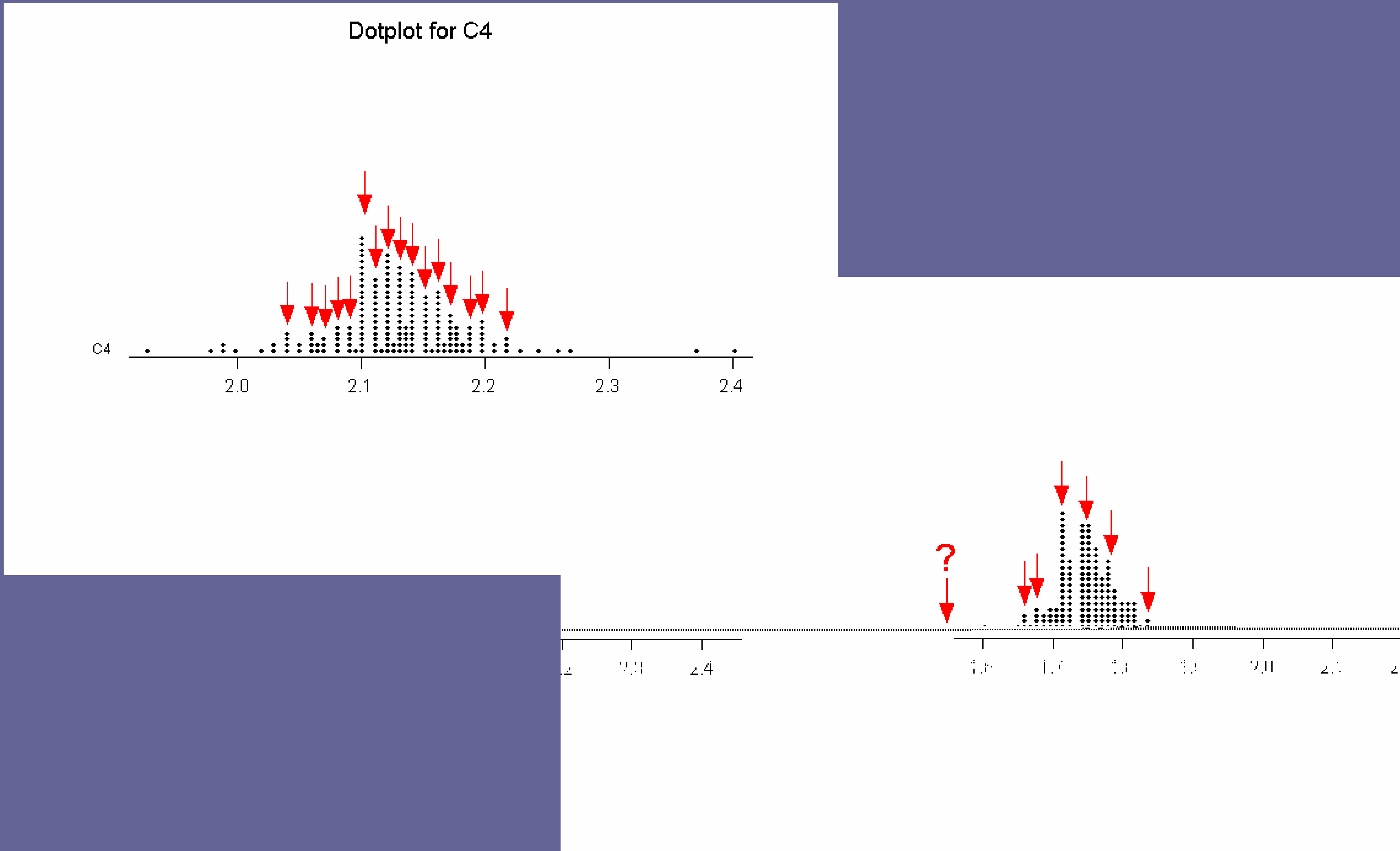


Possible use of a trimmed data set?



Can I use the mode?

How many modes? Where are they?



The normal kernel density for identifying a mode

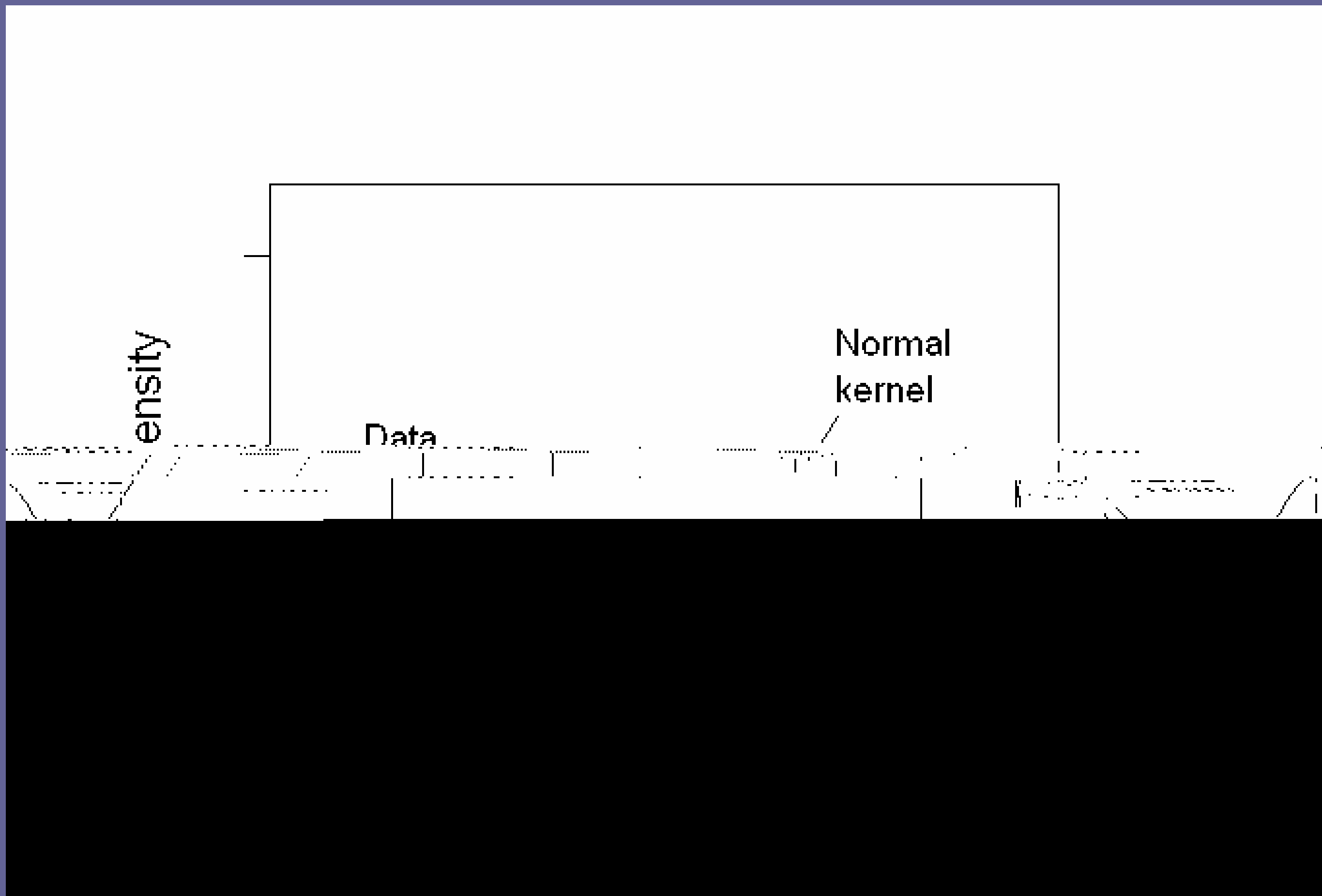
$$y = \frac{1}{nh} \sum_{i=1}^{n} \Phi\left(\frac{x - x_i}{h}\right)$$

where Φ is the standard normal density,

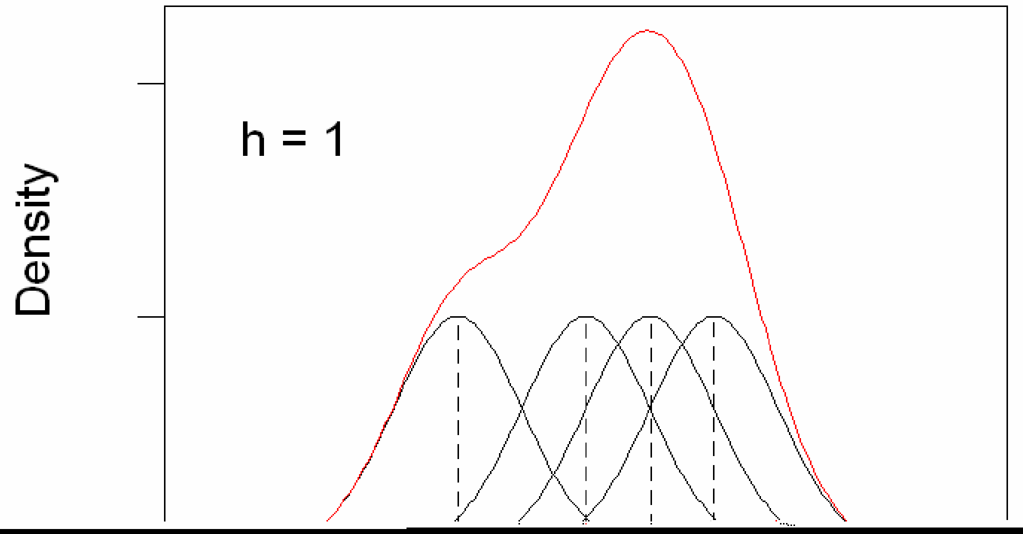
$$\Phi(a) = \frac{\exp(-a^2/2)}{\sqrt{2\pi}}$$

AMC Technical Brief No. 4

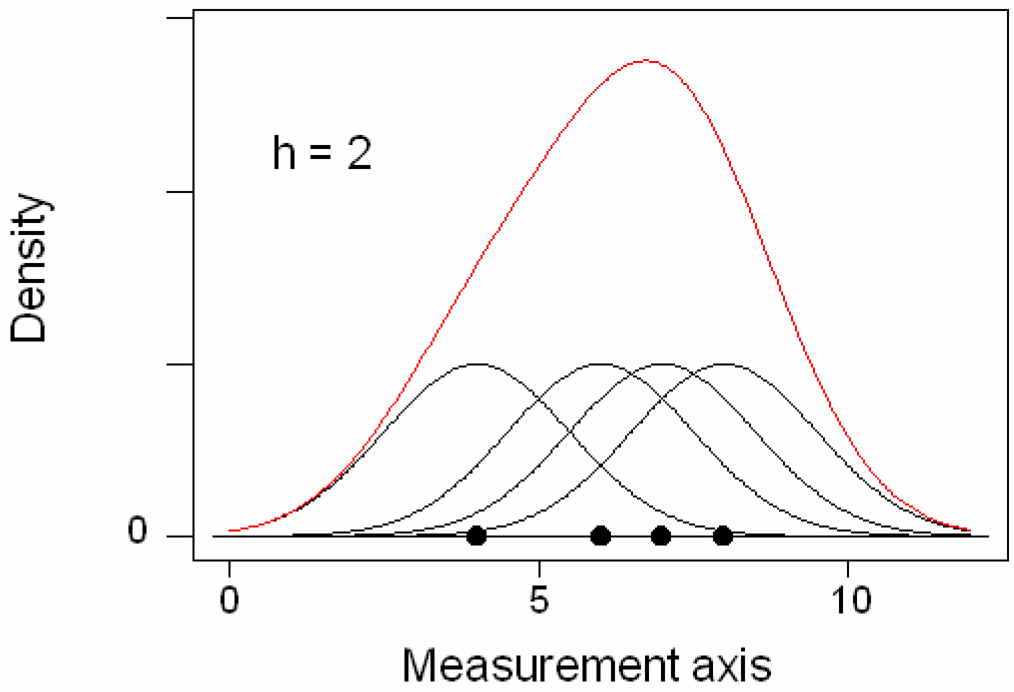
A normal kernel



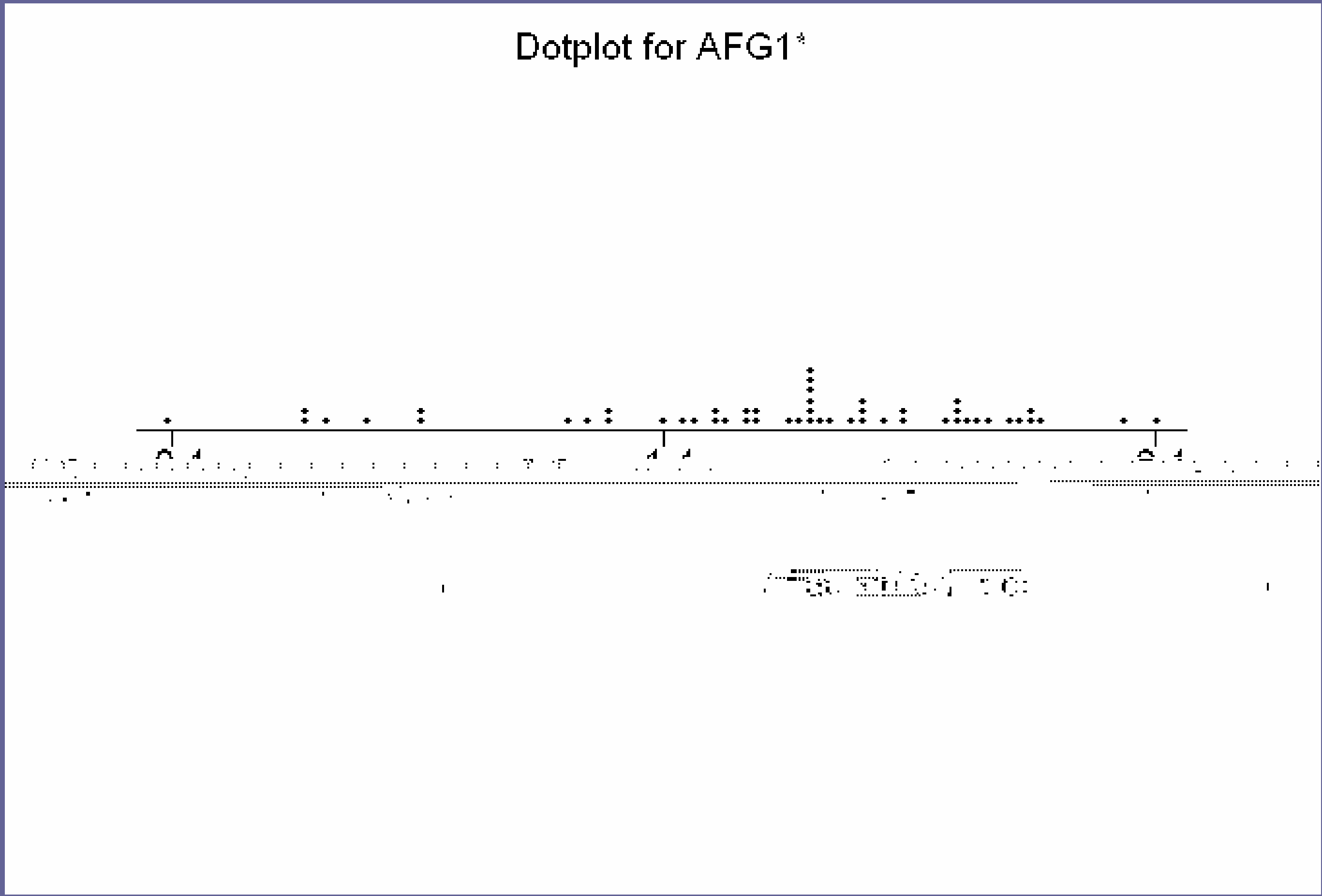
A kernel density



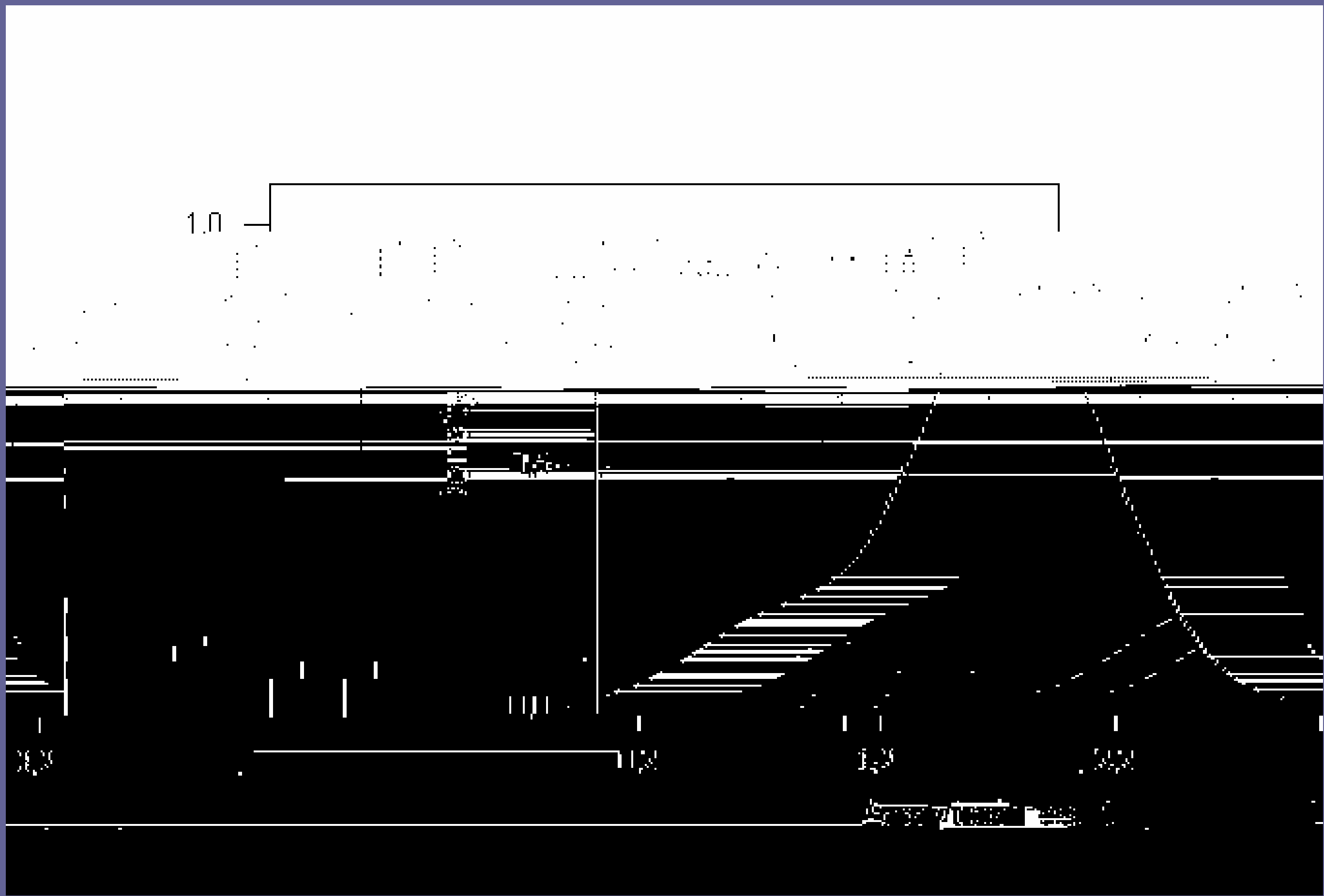
Another kernel density



Graphical representation of sample data



Kernel density of the aflatoxin data



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 –*

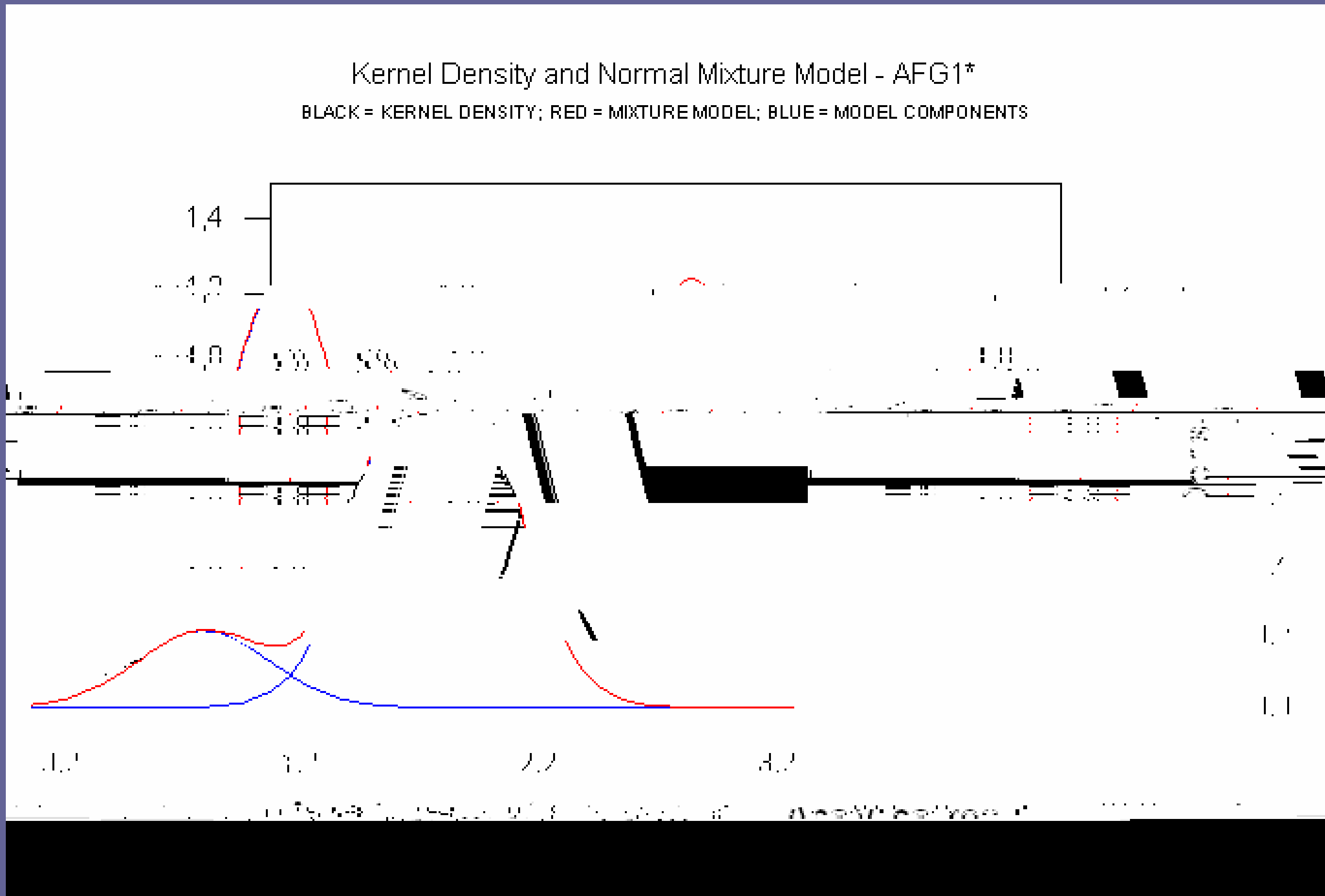
The normal mixture model

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

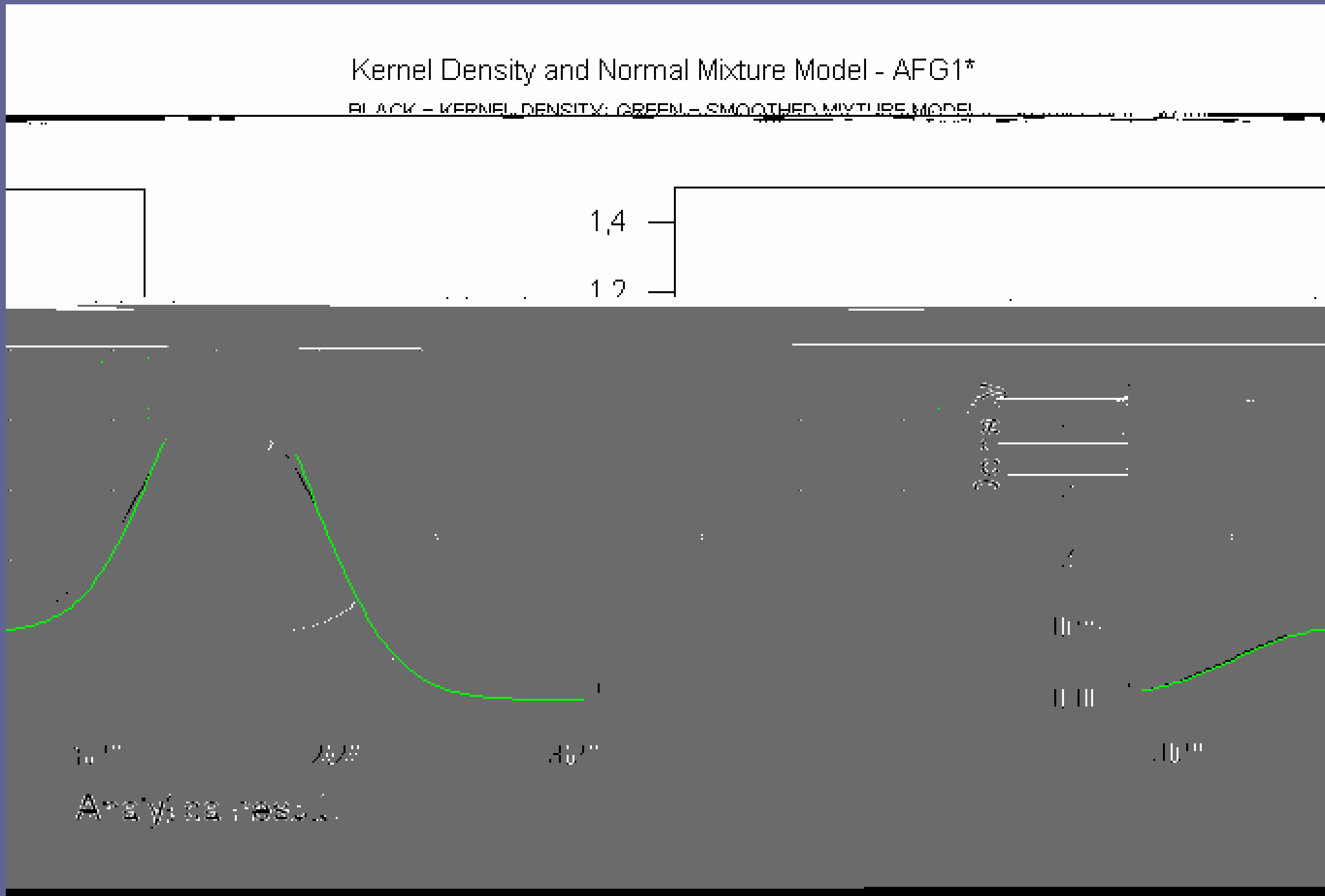
$$\frac{\quad}{\sqrt{\quad}}$$

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

Kernel density and fit of 2-component normal mixture model



Kernel density and variance-inflated mixture model



Useful 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

- Use z-

Homogeneity testing

- Comminute and mix bulk material.
- Split into distribution units.
- Select m>10 distribution units at random.
- Homogenise each one.
- Analyse 2 test portions from each in random order, with high precision, and conduct one-way analysis of variance on results.

Design for homogeneity testing

[redacted]

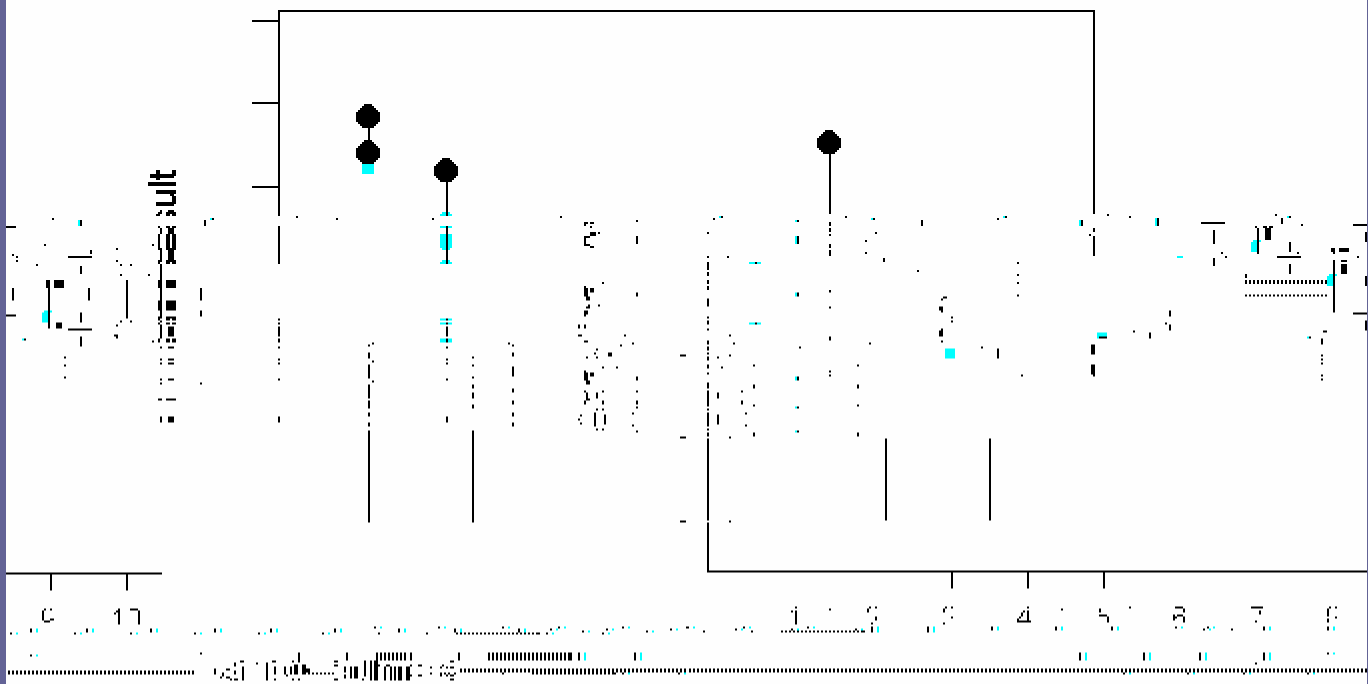
$$s_{an} \quad \sqrt{MSW}, \qquad s_{sam} \quad \sqrt{\frac{MSB \quad MSW}{2}}$$

Problems with simple ANOVA based on testing

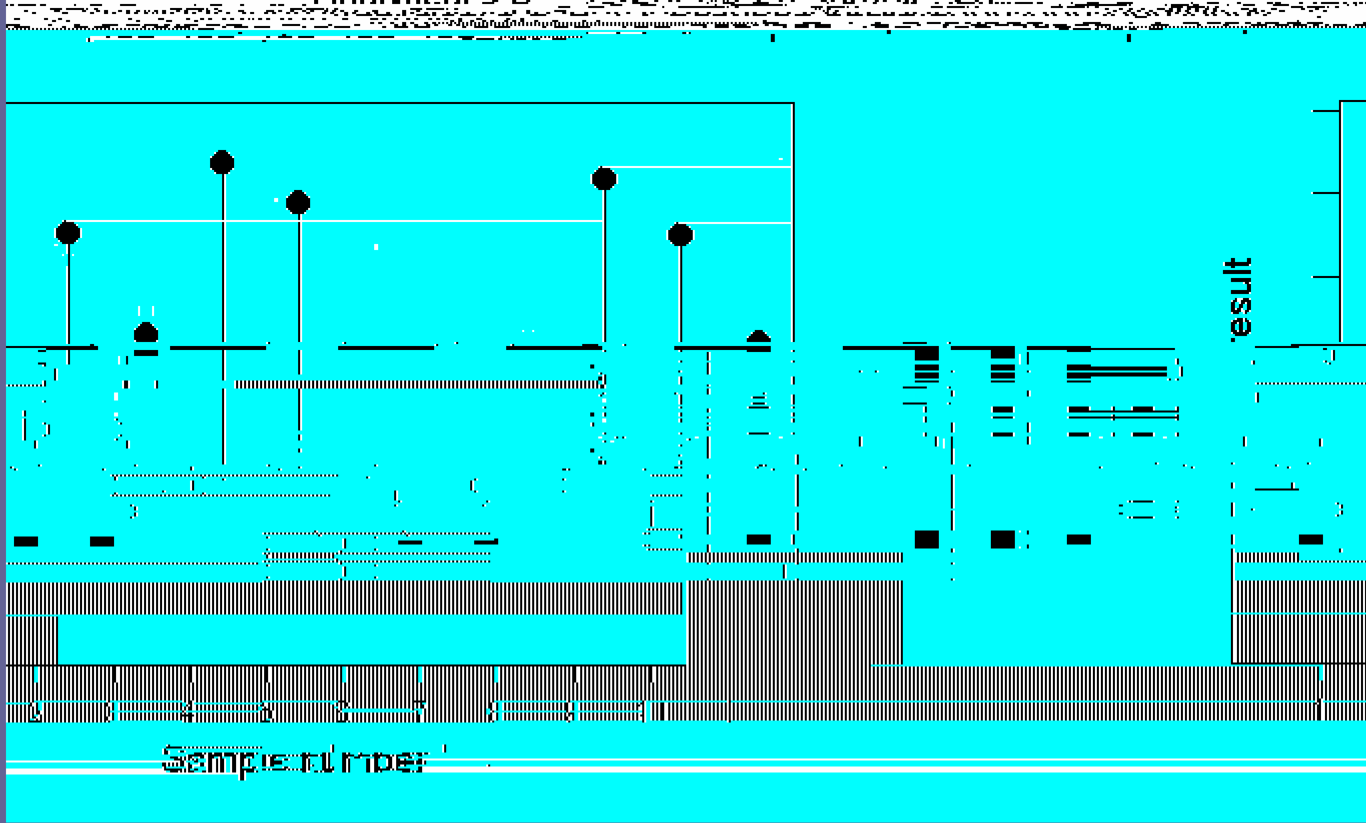
- Analytical precision too low—method

Unknowable true concentrations

Analytical s.d. = 0.5 X between-sample s.d.



Analytical s.d. = 2 X between-sample s.d.



- Material passes homogeneity test if

- Problems are:
  - s_{sam}

Fearn test

- Test $H_0 : \; {}^{2}_{sam} \qquad {}^{2}_{L}$ by rejecting when

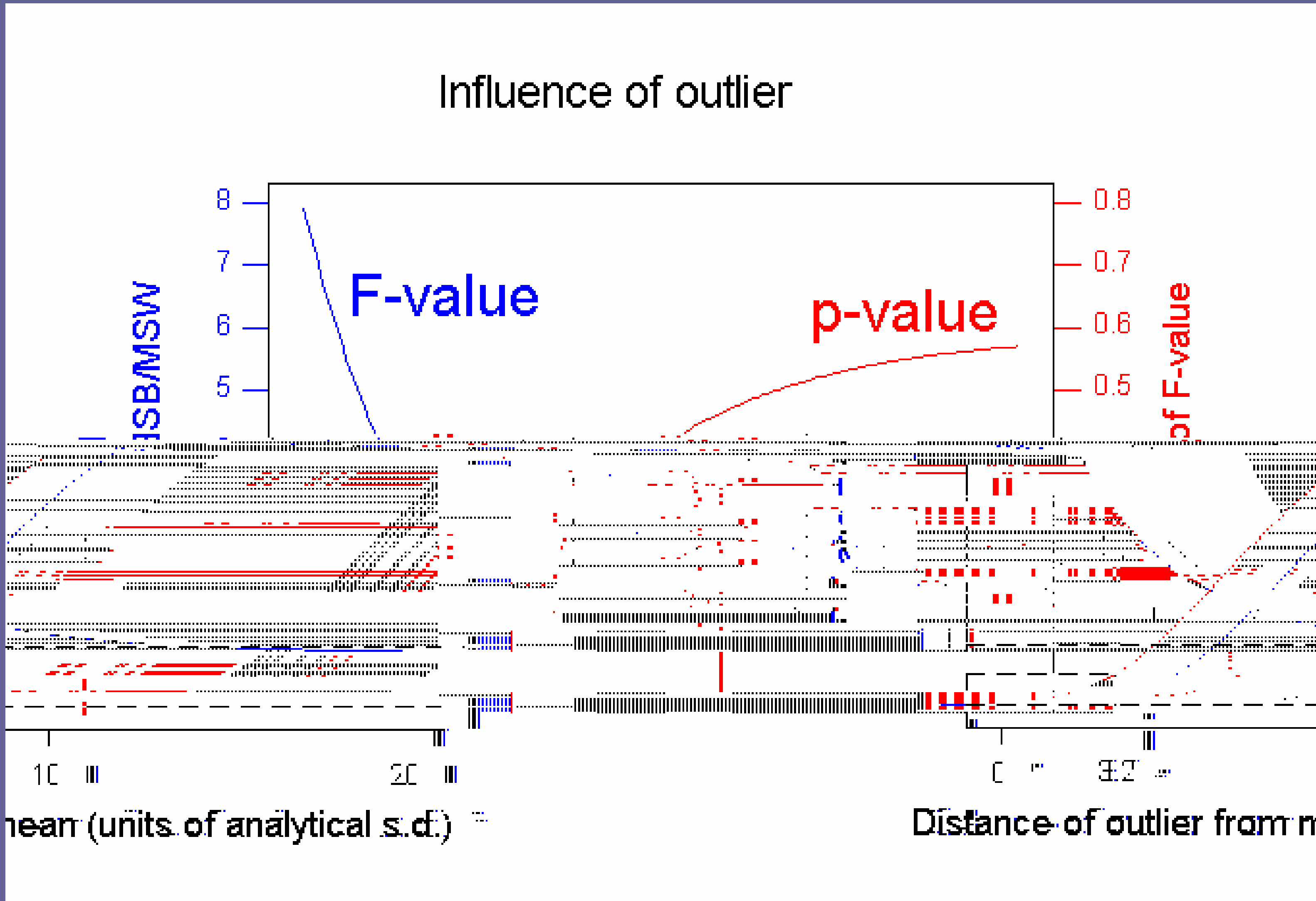
$$s^{2}_{sam} \qquad \frac{{}^{2}_{L} \quad {}^{2}_{m \;\; 1}}{m \quad 1} \qquad \frac{s^{2}_{an} \; F_{m \;\; 1, m} \qquad 1}{2}$$

Ref: *Analyst*, 2001, **127**, 1359-1364.

Problems with homogeneity data

- Problems with data are common: *e.g.*, no proper randomisation, insufficient precision, biases, trends, steps, insufficient significant figures recorded, outliers.
- Laboratories need detailed instructions.
- Data need careful scrutiny before statistics.
-

[redacted]


Influence of outlier
F-value
p-value
8
7
6
5
0.8
0.7
0.6
0.5
SB/MSW
of F-value
10
20
nean (units of analytical s.d.)
Distance of outlier from m

General references

- *The Harmonised Protocol* (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. International Standards Organisation, Geneva, 1997.