Setting standards
in analytical science

The 2006 IUPAC Harmonized Protocol for Proficiency Testing

M Thompson

S L R Ellison

R Wood

IUPAC Interdivisional Committee for the Harmonisation of Quality Control Systems

Setting standards
in analytical science

NEW!!
IUPAC Harmonized Protocol for Proficiency Testing

Slimmer, Fitter Scoring!!

Detects Multimodal sets!!

Cleaner Homogeneity tests!!

Scope of 2006 IUPAC protocol

- Only chemical analysis.
- Only results obtained on a fitness-for-purpose basis (*i.e.*, suitable for z-scoring with a pre-set value of σ_p).
- Only results on an interval scale or a ratio scale.
- Primarily scientific aspects
  - minimal administrative details
  - no criteria for assessment or accreditation of laboratories or PT schemes.

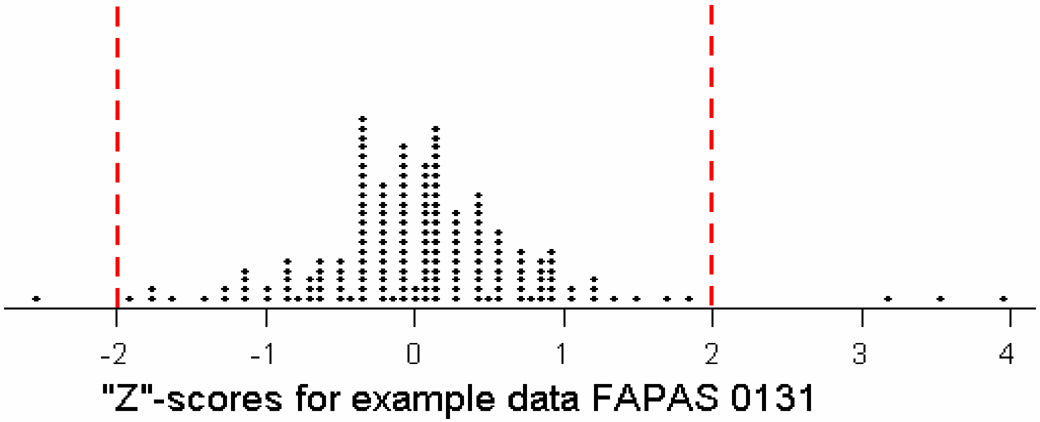
Properties of an ideal scoring method

- Adds value to raw results
  - Tells you more than just looking at raw data
- Easily understandable
  - e.g. based on the properties of the normal distribution.
- Has no arbitrary scaling transformation.
- Is transferable between different concentrations, analytes, matrices, and measurement principles.

A bad scoring method

Setting standards
in analytical science

$$z = \left(x - \overline{x}\right) / s$$



$$\overline{x} = 2.126$$

$$s = 0.077$$

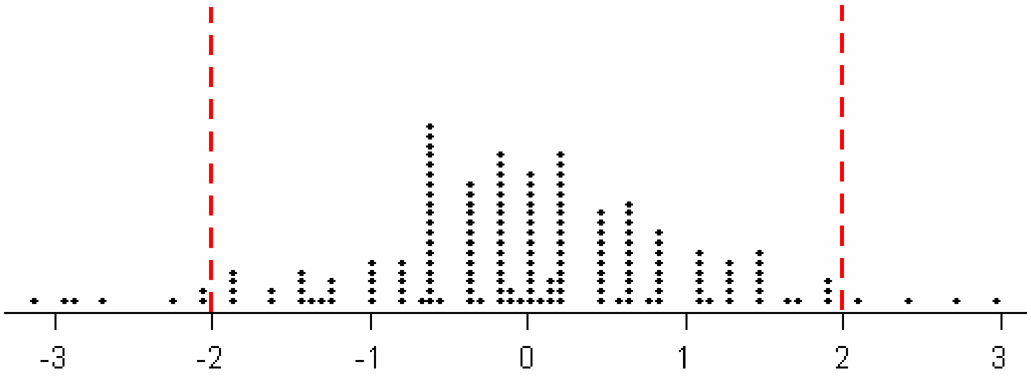
97% of scores in range $-2 < z < 2$

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

Another weak scoring method

Setting standards in analytical science

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



"Z"-score for example data FAPAS 0131

~91% of data within range -2 < z < 2

$$\hat{\mu}_{rob} = 2.128$$

$$\hat{\sigma}_{rob} = 0.048$$

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

2006 HP Scoring

Setting standards
in analytical science

- Focuses on the *z*-score

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

- ‘Fit-for-purpose’ scoring basis

$$\sigma_{\mathrm{p}} \equiv u_{\mathrm{ffp}}$$

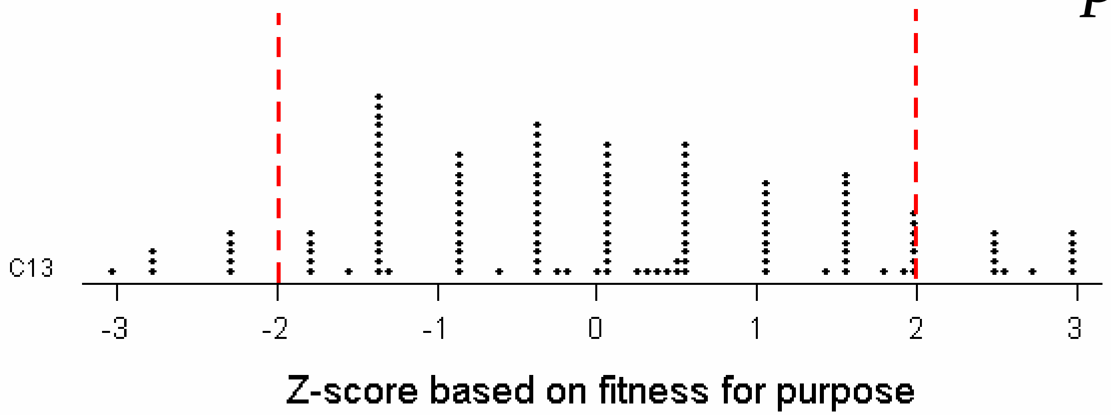
- Robustified against extreme values and informative about fitness for purpose.
- The protocol **is not restricted to consensus values**

“Fit-for-purpose” scoring: Example

- Set fitness for purpose criterion at RSD of 1%. This gives:

$$\sigma_p = 0.01 \times 2.1$$

$$= 0.021$$



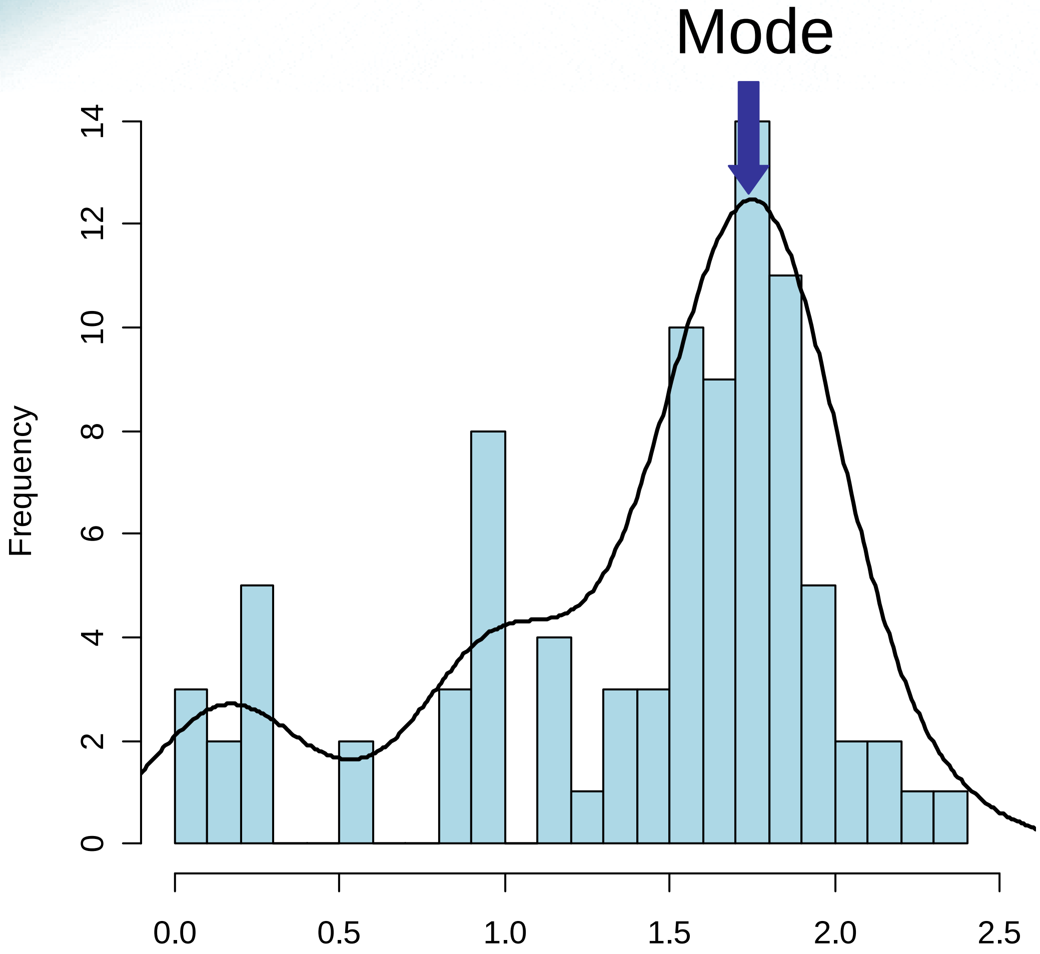
- About 78% within 0±2
  - ..for THIS data set with THIS criterion

Non-normal distributions

- Non-normal and multimodal distributions most commonly arise when the participants' results come from two or more inconsistent methods.
- Skews can arise as an artefact at low concentrations of analyte as a result of data recording (mal)practice.
- Sometimes skew can arise when the distribution is fundamentally non-normal
  - Example: GMO data expected to be approximately lognormal
  - **Transform before evaluation**

Handling Multimodal data

Setting standards in analytical science



FAPAS Arsenic data, round 0750

- Generate kernel density (h=0.75σ_p)
- Minor modes large
- Largest mode deemed ‘correct’*
- Use Kernel Density Mode

* If not, abandon scoring and investigate further

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.

Homogeneity testing in HP1/HP2: Procedure

- 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 ANOVA on results.

Homogeneity testing in HP1/HP2: Differences

- Rejects if

$$s_{sam} \leq 0.3\sigma_{p}$$

- Forbids outlier rejection

- Uses Thompson-Fearn test for “sufficient homogeneity”
- Requires (1) within-bottle outlier rejection

“Sufficient homogeneity” in HP1

Setting standards
in analytical science

- Material passes homogeneity test if

$$s_{sam} \leq 0.3\sigma_p$$

- Problems are:
  - s_{sam} may not be well estimated (9 degrees of freedom);
  - single-laboratory precision often close to $0.3\sigma_p$
  - too big a probability of rejecting satisfactory test material.

New protocol: Fearn-Thompson test

Setting standards in analytical science

- Test $H_0 : \sigma^2_{sam} < \sigma^2_{all} \;(\text{usually}\,0.3)$

- Reject when

$$s^2_{sam} > \frac{\sigma^2_{all}\chi^2_{m-1}}{m-1} + \frac{s^2_{an}\left(F_{m-1,m} - 1\right)}{2}$$

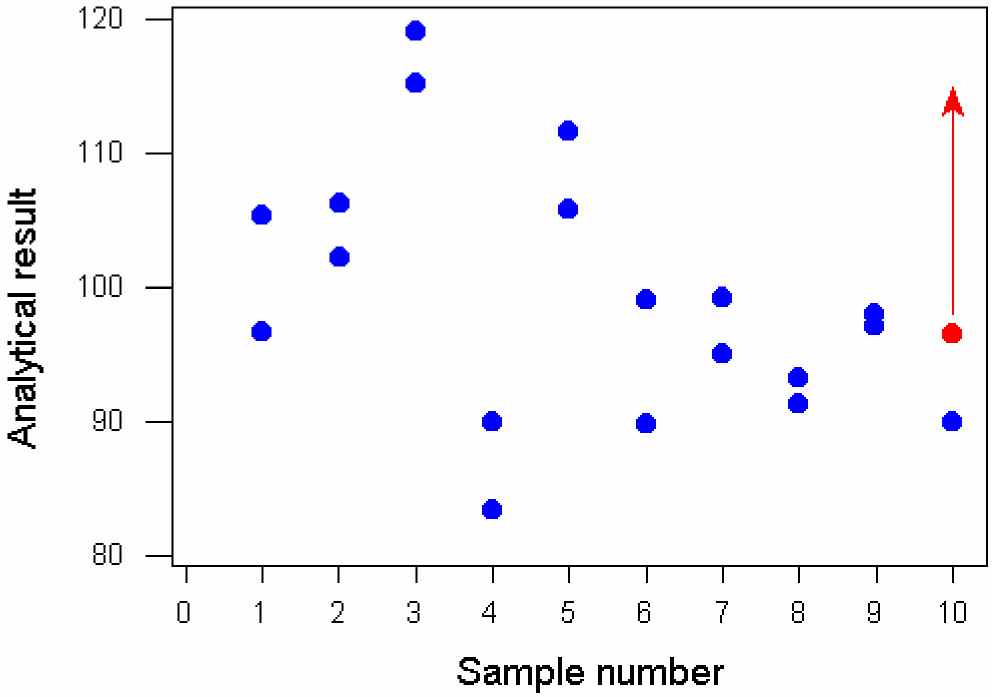
- Less likely to reject at random

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

Why use outlier rejection?

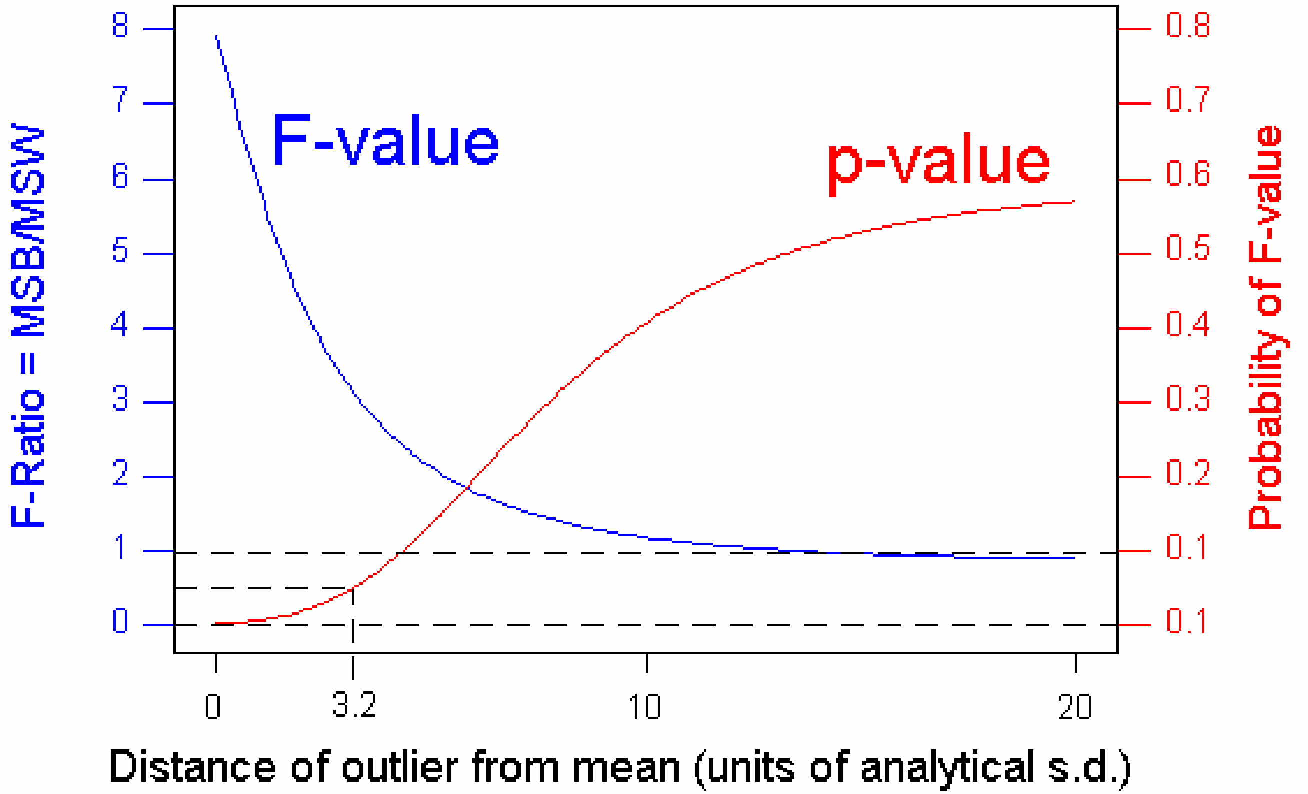
Setting standards
in analytical science

One-way ANOVA gives:
F = 9.5; p = 0.001



Why use outlier rejection?

Setting standards
in analytical science



Within-bottle outliers weaken the homogeneity test

Conclusion: New directions in the IUPAC protocol

LGC
Setting standards
in analytical science

- Stronger emphasis on fitness-for-purpose in scoring
- Clear acceptance of continued use of consensus values
  - with advice on implementation
- Testing for statistical evidence of insufficient homogeneity instead of fixed value
- Does not recommend that the organiser provide scores based on participant uncertainties
  - DOES control uncertainties in assigned value
  - Provides methods for participants to assess their own uncertainty and fitness for purpose