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

“Target value” or

“standard deviation for proficiency”

\[ z = \frac{x - x_A}{\sigma_p} \]
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 (\textit{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

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

\[
x^T = [x_1 \ x_2 \ \Lambda \ x_n]
\]

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

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

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

If not converged, $p = p + 1$
When can I safely use robust estimates?

Measurement axis

- Skewed
- Bimodal
- Heavy-tailed

- ✓
- ✗
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 \( \Phi \) 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.

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

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

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

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

- The M-step

\[
\hat{p}_j = \frac{\sum_{i=1}^{n} \hat{P}(j|y_i)}{n} \\
\hat{\mu}_j = \frac{\sum_{i=1}^{n} y_i \hat{P}(j|y_i)}{\sum_{i=1}^{n} \hat{P}(j|y_i)} \\
\hat{\sigma}^2 = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} (y_i - \hat{\mu}_j)^2 \hat{P}(j|y_i)}{\sum_{i=1}^{n} \hat{P}(j|y_i)}
\]

- The E-step

\[
\hat{P}(j|y_i) = \hat{p}_j f_j(y_i) / \sum_{j=1}^{m} \hat{p}_j f_j(y_i)
\]
Example datasets
Example dataset 1

Nitrogen in canned meat

Result, % by mass

Laboratory ID

z = -2

z = 0

z = 2
Nitrogen in canned meat

Result, % by mass
Number of modes vs smoothing factor $h$
Nitrogen in canned meat

BLACK = KERNEL DENSITY
RED = MIXTURE MODEL

COMPONENT 1: MEAN = 2.0120; SD = 0.0747; p = 1
Bootstrapped kernel density plots

Result, % by mass

Density
Statistics: dataset 1

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{\mu}$</th>
<th>$\hat{\sigma}$</th>
<th>$se(\hat{\mu})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust</td>
<td>2.912</td>
<td>0.056</td>
<td>0.0056</td>
</tr>
<tr>
<td>Kernel density mode</td>
<td>2.912</td>
<td>-</td>
<td>0.0056</td>
</tr>
<tr>
<td>Mixture model</td>
<td>2.913</td>
<td>0.075</td>
<td>0.0075</td>
</tr>
</tbody>
</table>
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

Result, % by mass
Polyunsaturated fatty acids

Result, % by mass
Polyunsaturated fatty acids

Horwitz standard deviation

Result, % by mass

Possible bimodal distribution?
Polyunsaturated fatty acids

BLACK = KERNEL DENSITY; RED = MIXTURE MODEL; BLUE = MODEL COMPONENTS

COMPONENT 1: MEAN = 30.318; SD = 0.370; p = 0.69
COMPONENT 2: MEAN = 40.535; SD = 0.291; p = 0.32

Result, % by mass
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

GeoPT Round 20. Test material: silicate rock
COMPONENT 1: MEAN = 5.930; SD = 0.501; p = 0.620
COMPONENT 2: MEAN = 9.720; SD = 1.103; p = 0.242
COMPONENT 3: MEAN = 15.235; SD = 1.273; p = 0.138

GeoPT-20--Ba

Density

Mixture model

Kernel density

Analytical result
Choice of value for $\sigma_p$

- Robust standard deviation of participants’ results in round?
- From perception of how well similar methods perform?
- Legislation?
- Other?
Self-referential scoring

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

- 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.

A score that meets all of the criteria

- If we now define a z-score thus:
  \[ z = \frac{x - \hat{\mu}_{rob}}{\sigma_p} \]  
  where \( \sigma_p \equiv u_f \)
  
  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**
  AMC Technical Brief No. 23, 2006. www/rsc.org/amc

• **Kernel densities**
  B W Silverman, *Density estimation for statistics and data analysis.*
  AMC Technical Brief, no. 4, 2001 www/rsc.org/amc

• **The bootstrap**
  B Efron and R J Tibshirani, *An introduction to the bootstrap.*
  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)
General references


