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
Organisation of a proficiency test

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.

\[ \mu \pm 3\sigma \]

\[ \mu \pm 2\sigma \]

\[ \mu \pm \sigma \]

\[ \mu \]

\[ \mu - \sigma \]

\[ \mu - 2\sigma \]

\[ \mu - 3\sigma \]

\[ \sim 95\% \]

\[ \sim 99.7\% \]

BUT…

We do not make any assumptions about the actual data.
Example dataset A

- Determination of protein nitrogen in a meat product.
A weak scoring method

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

\[ z = \left( x - \bar{x} \right) / s \]

\[ \bar{x} = 2.126 \]

\[ s = 0.077 \]
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.
Can I use robust estimates?

Skewed \(\times\)

Bimodal \(\times\)

Heavy-tailed \(\checkmark\)
Huber’s H15

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

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

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

If not converged, $p = p + 1$
References: robust statistics

- Analytical Methods Committee, *Analyst*, 1989, **114**, 1489
- AMC Technical Brief No 6, 2001 (download from www/rsc.org/amc)
Is that enough?

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

\[
z = \frac{x - \mu_{rob}}{\hat{\sigma}_{rob}}\]

\[
\hat{\mu}_{rob} = 2.128
\]

\[
\hat{\sigma}_{rob} = 0.048
\]
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.

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 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:

$$\sigma_p = 0.01 \times 2.1 = 0.021$$

Data from FAPAS round 0131

Z-score based on fitness for purpose

78% of z-scores within range $-2 < z < 2$
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

\[
u(x_a) = \hat{\sigma}_{rob} / \sqrt{n}
\]
When can I use robust estimates?

Measurement axis

- Skewed (X)
- Bimodal (X)
- Heavy-tailed (√)
Skewed distributions

- Skews can arise when the participants’ results come from two or more inconsistent methods.
- They can also arise as an artefact at low concentrations of analyte as a result of data recording practice.
- Rarely, skews can arise when the distribution is truly lognormal.
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 \( \Phi \) is the standard normal density,

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

AMC Technical Brief No. 4
A normal kernel

![Graph showing a normal kernel with data points and a measurement axis.](image-url)
A kernel density
Another kernel density

$h = 2$
Graphical representation of sample data

Dotplot for AFG1

Aflatoxin G1, ppb
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.

The normal mixture model

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

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

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

- The M-step

\[ \hat{P}_j = \frac{\sum_{i=1}^{n} \hat{P}(j | y_i)}{n} \]

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

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

- The E-step

\[ \hat{P}(j | y_i) = \frac{\hat{p}_j f_j(y_i)}{\sum_{j=1}^{m} \hat{p}_j f_j(y_i)} \]
Kernel density and fit of 2-component normal mixture model

**Kernel Density and Normal Mixture Model - AFG1**
- **Black** = Kernel Density
- **Red** = Mixture Model
- **Blue** = Model Components

**Analytical Result**

Component 1: MEAN = 0.7901, SD = 0.2753; p = 0.192
Component 2: MEAN = 1.7774, SD = 0.2753; p = 0.848
Kernel density and variance-inflated mixture model

Kernel Density and Normal Mixture Model - AFG1*

BLACK = KERNEL DENSITY; GREEN = SMOOTHEDE MIXTURE MODEL

Density

0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4

0.2 1.2 2.2 3.2

Analytical result
Useful References

• **Mixture models**
  AMC Technical Brief No. 23, 2006. www/rsc.org/amc

• **Kernel densities**
  AMC Technical Brief, no. 4, 2001 www/rsc.org/amc

• **The bootstrap**
  AMC Technical Brief, No. 8, 2001 www/rsc.org/amc
Conclusions—scoring

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

\[ s_{an} = \sqrt{MSW}, \quad s_{sam} = \sqrt{\frac{MSB - MSW}{2}} \]
Problems with simple ANOVA based on testing

\[ H_0 : \sigma_{sam} = 0 \]

- Analytical precision too low—method cannot detect consequential degree of heterogeneity.
- Analytical precision too high—method finds significant degree of heterogeneity that may not be consequential.

(Everything is heterogeneous!)
Analytical s.d. = 0.5 X between-sample s.d.
Analytical s.d. = 2 X between-sample s.d.
“Sufficient homogeneity”: original definition

• Material passes homogeneity test if

\[ s_{sam} \leq \sigma_L = 0.3\sigma_p \]

• Problems are:
  – \( s_{sam} \) may not be well estimated;
  – too big a probability of rejecting satisfactory test material.
Fearn test

- Test \( H_0 : \sigma^2_{sam} < \sigma^2_L \) by rejecting when

\[
s_{sam}^2 > \frac{\sigma^2_L \chi^2_{m-1}}{m-1} + \frac{s^2_{an}(F_{m-1,m} - 1)}{2}
\]

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.
- HP1 is incorrect in saying that all outlying data should be retained.
One-way ANOVA gives:
F = 9.5;  p = 0.001
Influence of outlier

F-value

p-value

F-Ratio = MSB/MSW

Distance of outlier from mean (units of analytical s.d.)

Probability of F-value
General references

• *The Harmonised Protocol* (revised)
  M Thompson, S L R Ellison and R Wood

• R E Lawn, M Thompson and R F Walker,
  *Proficiency testing in analytical chemistry*. The

• ISO Guide 43. International Standards