Consensus Building: Statistical Treatment of Multiple Results of the Same Measurand

Amanda Koepke

Joint work with Antonio Possolo, NIST Fellow & Chief Statistician

March 2019

NIST

Outline

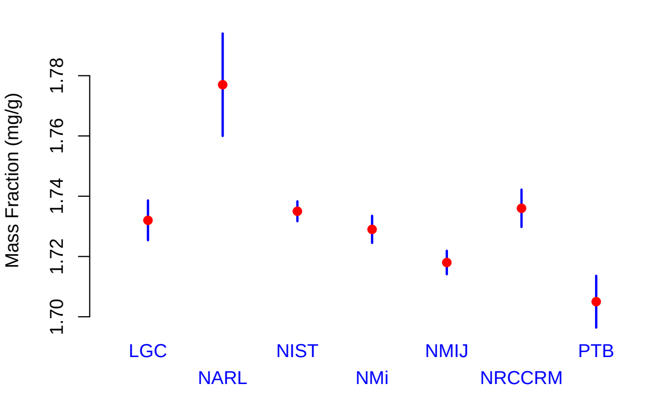
1. Definitions
2. Principles
3. Statistical Methods
4. Degrees of Equivalence

Consensus building

- Combine measurement results into consensus estimate.

Consensus building

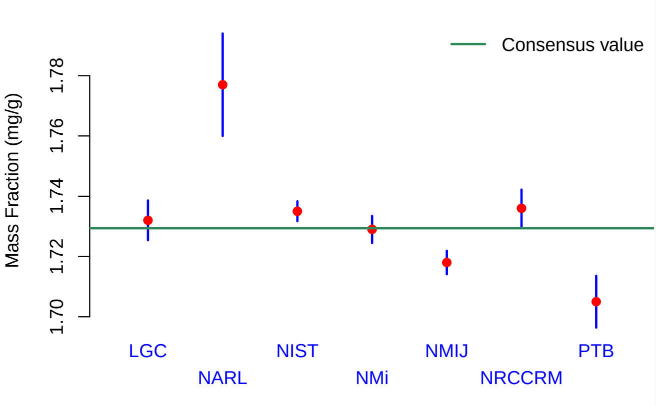
- Combine measurement results into consensus estimate.



CCQM-K6: Cholesterol in Human Serum

Consensus building

- Combine measurement results into consensus estimate.



CCQM-K6: Cholesterol in Human Serum

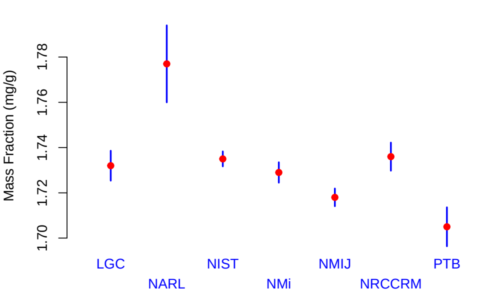
Consensus building

- Combine measurement results into consensus estimate.
- Qualify consensus estimate with evaluation of measurement uncertainty that captures
  - Stated uncertainties associated with individual measured values

Consensus building

- Combine measurement results into consensus estimate.
- Qualify consensus estimate with evaluation of measurement uncertainty that captures
  - Stated uncertainties associated with individual measured values
  - Dark uncertainty (Thompson & Ellison, 2011)

Dark uncertainty

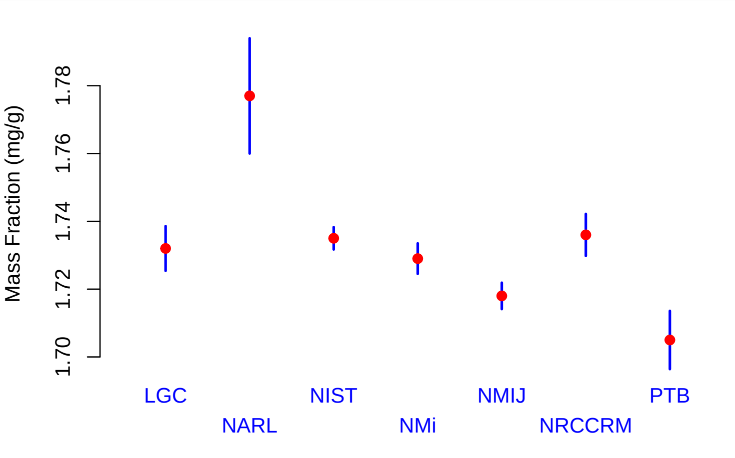


- Analogy to ‘dark matter’
- Uncovered when measured values are intercompared
- Unexpectedly large dispersion of values among the labs

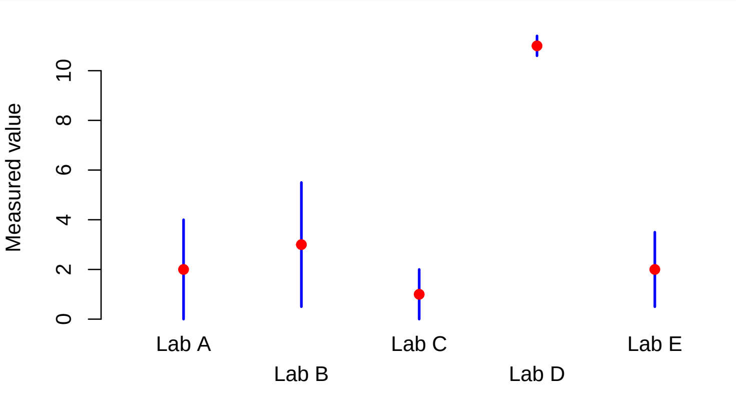
Principles

Principles

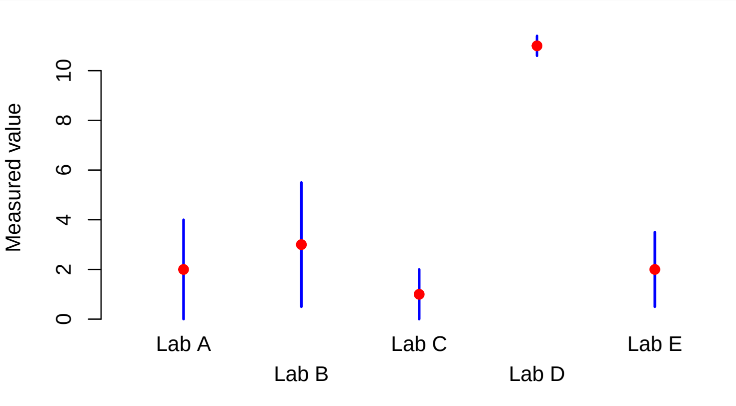
P1: The statistical model used for analysis should be able to detect, evaluate, and propagate dark uncertainty.



Principles



Principles

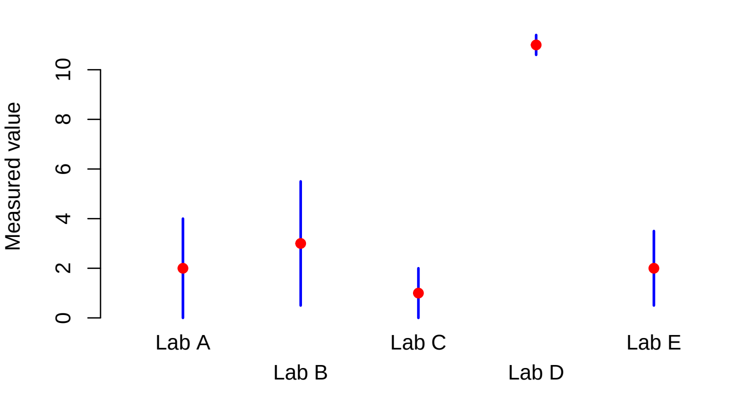


P2: No measurement result should be set aside except for substantive, documented cause.

- Graphical and statistical detection of anomalous results are useful screening tools, but should be advisory

Principles

P3: No measured value should dominate consensus value simply because it has a small uncertainty.



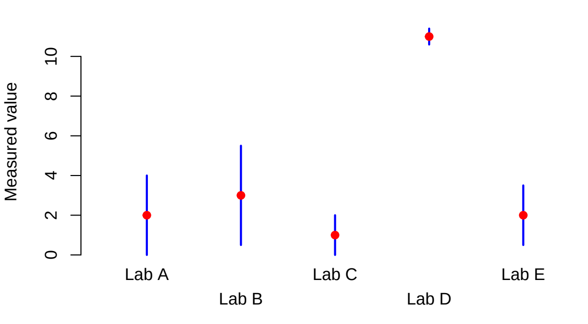
Principles

P3: No measured value should dominate consensus value simply because it has a small uncertainty.

- Weighted mean

$$\widehat{\mu} = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j}$$

$$w_j = \frac{1}{u_j^2}$$



Principles

P3: No measured value should dominate consensus value simply because it has a small uncertainty.

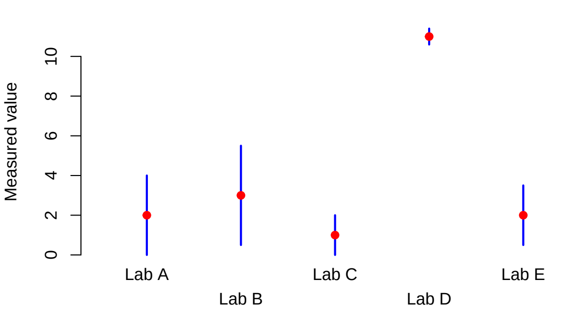
- Weighted mean

$$\widehat{\mu} = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j}$$

$$w_j = \frac{1}{u_j^2}$$

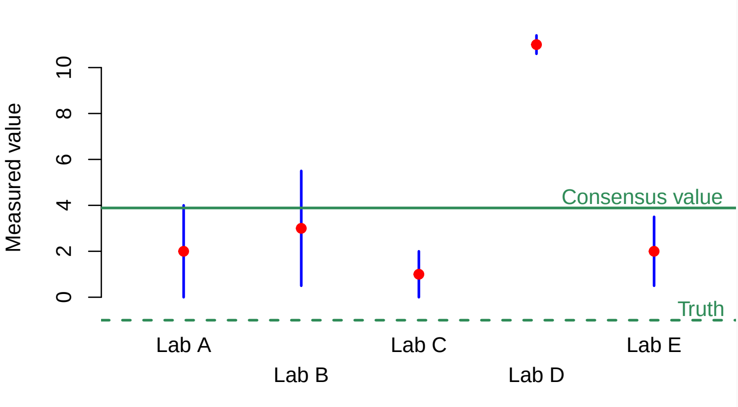
- Instead use:

$$w_j = \frac{1}{u_j^2 + \tau^2}$$



Principles

P4: Participating laboratories/methods should be selected and characterized sufficiently well to warrant belief that measured values are roughly centered at the true value of measurand.



Summary of Principles

- Dark uncertainty
- Outliers
- Belief that laboratories/methods are selected and characterized well

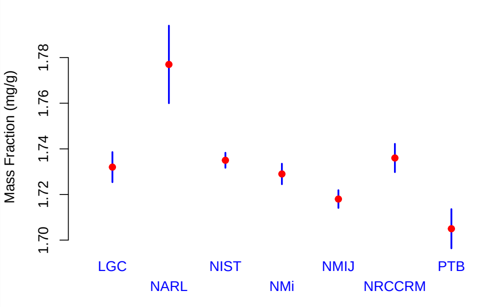
Summary of Principles

- Dark uncertainty
- Outliers
- Belief that laboratories/methods are selected and characterized well

"The willingness of the participants in an interlaboratory study to engage in an intercomparison should include a tacit agreement to abide by the resulting findings, which create the opportunity for collective learning and provide a stimulus for improving measurement quality." – Koepke et al (2017), *Metrologia*

Methods

CCQM-K6: Cholesterol in Human Serum

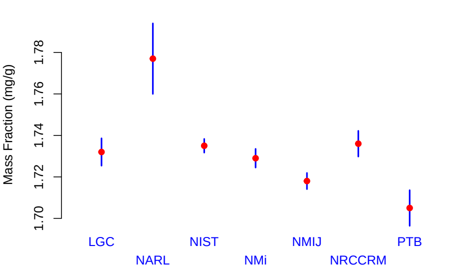


Lab	x	u	ν
LGC	1.732	0.0066	60
NARL	1.777	0.0170	11.3
NIST	1.735	0.0033	13.5
NMi	1.729	0.0045	60
NMIJ	1.718	0.0039	27
NRCCRM	1.736	0.0062	7.4
PTB	1.705	0.0086	314

Random effects model

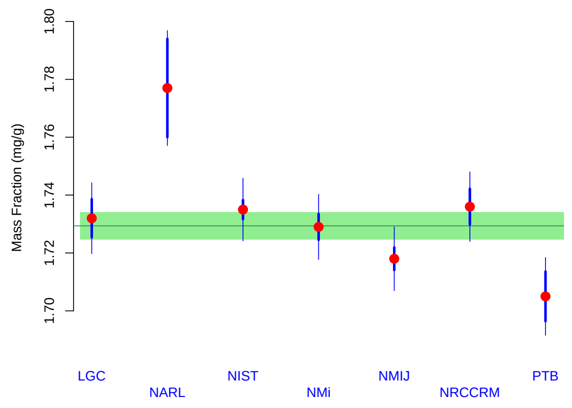
$$x_j = \mu + \lambda_j + \varepsilon_j \text{ for } j = 1, \ldots, n$$

- x_j Value measured by lab j
- μ Measurand
- λ_j Effect of lab j
  $\lambda_j \sim \mathrm{N}(0, \tau^2)$
- ε_j Measurement error for lab j
  $\varepsilon_j \sim \mathrm{N}(0, \sigma_j^2)$



DerSimonian-Laird procedure

- $\widehat{\mu} = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j}$
  $w_j = \frac{1}{\tau^2 + \sigma_j^2}$
- $u_{\mathrm{DL}}(\mu) = \sqrt{\frac{1}{\sum_{j=1}^{n} w_j}}$



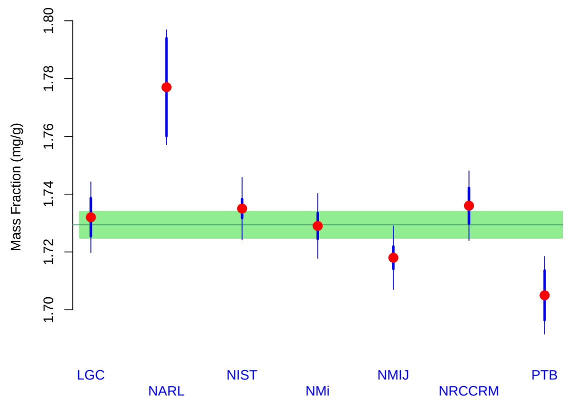
DerSimonian-Laird procedure

- $\widehat{\mu} = \frac{\sum_{j=1}^{n} w_j x_j}{\sum_{j=1}^{n} w_j}$

  $w_j = \frac{1}{\tau^2 + \sigma_j^2}$

- $u_{\mathrm{DL}}(\mu) = \sqrt{\frac{1}{\sum_{j=1}^{n} w_j}}$

- $\{\sigma_j\}$ and τ are unknown
  - $\widehat{\sigma}_j = u_j$
  - $\widehat{\tau}_{\mathrm{DL}} = \max\{0, \widehat{\tau}_{\mathrm{M}}\}$



- $\widehat{\tau}_{\mathrm{M}}^2 = \frac{Q - n + 1}{\sum_{j=1}^{n} u_j^{-2} - \sum_{j=1}^{n} u_j^{-4} / \sum_{j=1}^{n} u_j^{-2}}$, where $Q = \sum u_j^{-2}(x_j - \widehat{\mu})^2$

Bayesian approach

$$p(\theta|y) \propto p(\theta) \times p(y|\theta) = \text{Prior} \times \text{Likelihood}$$

Bayesian approach

$$p(\theta|y) \propto p(\theta) \times p(y|\theta) = \text{Prior} \times \text{Likelihood}$$

- Same random effects model: $x_j = \mu + \lambda_j + \varepsilon_j$ for $j = 1, \ldots, n$

Bayesian approach

$$p(\theta|y) \propto p(\theta) \times p(y|\theta) = \text{Prior} \times \text{Likelihood}$$

- Same random effects model: $x_j = \mu + \lambda_j + \varepsilon_j$ for $j = 1, \ldots, n$
- Can easily incorporate:
  - Uncertainty in σ_j^2
    - $\nu_j u_j^2 / \sigma_j^2 \sim \chi^2(\nu_j)$
  - Uncertainty in τ estimate
  - Any other prior knowledge

Bayesian analysis

Posterior density: $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}) \propto p(\boldsymbol{\theta}) \times p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}|\boldsymbol{\theta})$

- Unknown: $\boldsymbol{\theta} = (\mu, \tau, \boldsymbol{\lambda}, \boldsymbol{\sigma})$

Bayesian analysis

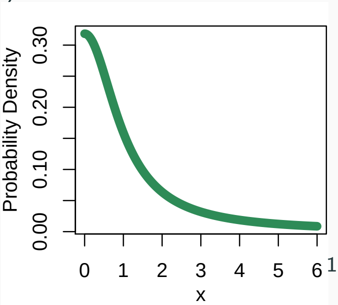
Posterior density: $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}) \propto p(\boldsymbol{\theta}) \times p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}|\boldsymbol{\theta})$

- Unknown: $\boldsymbol{\theta} = (\mu, \tau, \boldsymbol{\lambda}, \boldsymbol{\sigma})$
- Prior: $p(\boldsymbol{\theta}) = p(\mu)p(\tau)\prod_{i=1}^{n}[p(\lambda_j|\tau)p(\sigma_j)]$
    - $\mu \sim \text{Normal}(0, 10^5)$
    - $\tau \sim \text{half-Cauchy(scale=}\textit{Large})$
    - $\sigma_j \sim \text{half-Cauchy(scale=}\textit{Large})$
    - $\lambda_j|\tau \sim \text{Normal}(0, \tau^2)$

Bayesian analysis

Posterior density: $p(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{u},\boldsymbol{\nu}) \propto p(\boldsymbol{\theta}) \times p(\boldsymbol{x},\boldsymbol{u},\boldsymbol{\nu}|\boldsymbol{\theta})$

- Unknown: $\boldsymbol{\theta} = (\mu, \tau, \boldsymbol{\lambda}, \boldsymbol{\sigma})$
- Prior: $p(\boldsymbol{\theta}) = p(\mu)p(\tau)\prod_{i=1}^{n}[p(\lambda_j|\tau)p(\sigma_j)]$
  - $\mu \sim \text{Normal}(0, 10^5)$
  - $\tau \sim \text{half-Cauchy}(\text{scale}=\textit{Large})$
  - $\sigma_j \sim \text{half-Cauchy}(\text{scale}=\textit{Large})$
  - $\lambda_j|\tau \sim \text{Normal}(0, \tau^2)$



Bayesian analysis

Posterior density: $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}) \propto p(\boldsymbol{\theta}) \times p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}|\boldsymbol{\theta})$

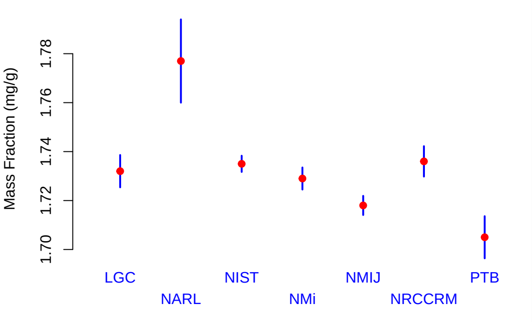
- Unknown: $\boldsymbol{\theta} = (\mu, \tau, \boldsymbol{\lambda}, \boldsymbol{\sigma})$
- Prior: $p(\boldsymbol{\theta}) = p(\mu)p(\tau)\prod_{i=1}^{n}[p(\lambda_j|\tau)p(\sigma_j)]$
  - $\mu \sim \text{Normal}(0, 10^5)$
  - $\tau \sim \text{half-Cauchy(scale=}\textit{Large})$
  - $\sigma_j \sim \text{half-Cauchy(scale=}\textit{Large})$
  - $\lambda_j|\tau \sim \text{Normal}(0, \tau^2)$
- Likelihood: $p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{u}, \boldsymbol{\nu}|\boldsymbol{\theta})$
  - $x_j|\mu, \lambda_j, \sigma_j \sim \text{Normal}(\mu + \lambda_j, \sigma_j^2)$
  - $\frac{\nu_j u_j^2}{\sigma_j^2}|\sigma_j \sim \chi^2(\nu_j)$

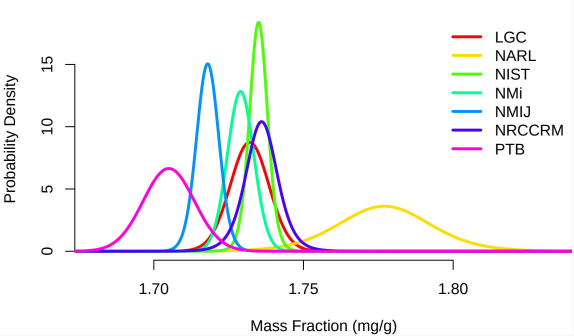
Bayesian analysis

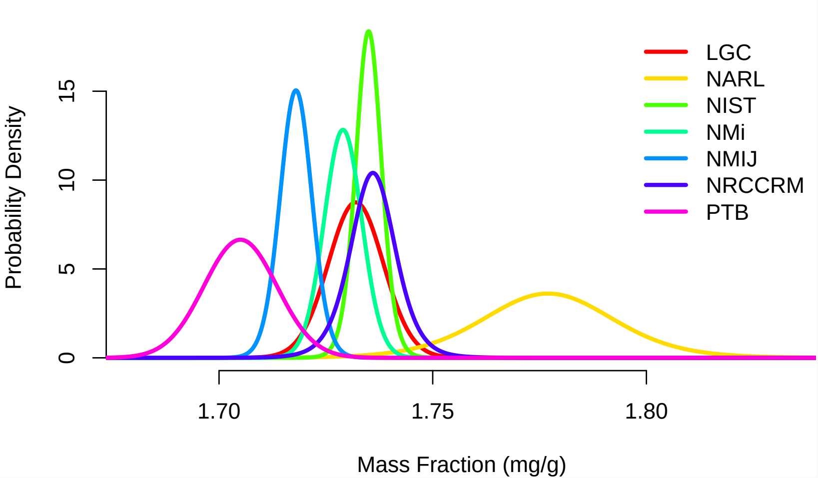
Posterior density: $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}) \propto p(\boldsymbol{\theta}) \times p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}|\boldsymbol{\theta})$

- Unknown: $\boldsymbol{\theta} = (\mu, \tau, \boldsymbol{\lambda}, \boldsymbol{\sigma})$
- Prior: $p(\boldsymbol{\theta}) = p(\mu)p(\tau)\prod_{i=1}^{n}[p(\lambda_j|\tau)p(\sigma_j)]$
    - $\mu \sim \text{Normal}(0, 10^5)$
    - $\tau \sim$ half-Cauchy(scale=*Large*)
    - $\sigma_j \sim$ half-Cauchy(scale=*Large*)
    - $\lambda_j|\tau \sim \text{Normal}(0, \tau^2)$
- Likelihood: $p(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\nu}|\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{u}, \boldsymbol{\nu}|\boldsymbol{\theta})$

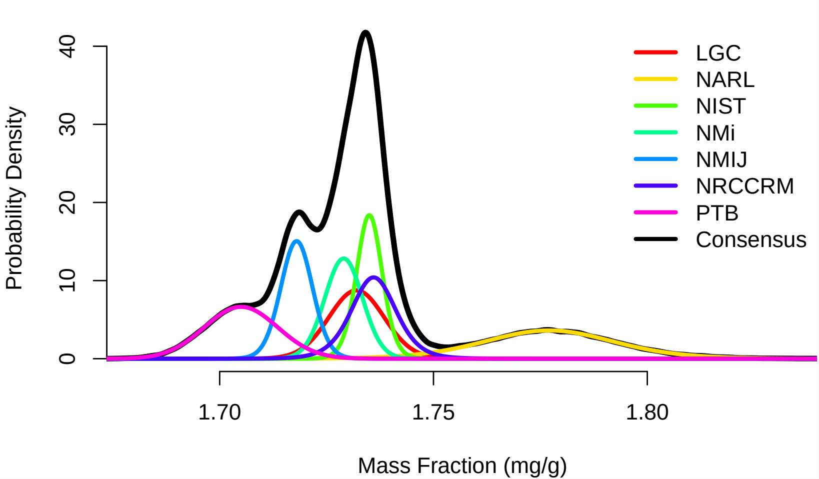
Simulate from the posterior distribution using Markov chain Monte Carlo.


Mass Fraction (mg/g)
1.70
1.72
1.74
1.76
1.78
LGC
NARL
NIST
NMi
NMIJ
NRCCRM
PTB

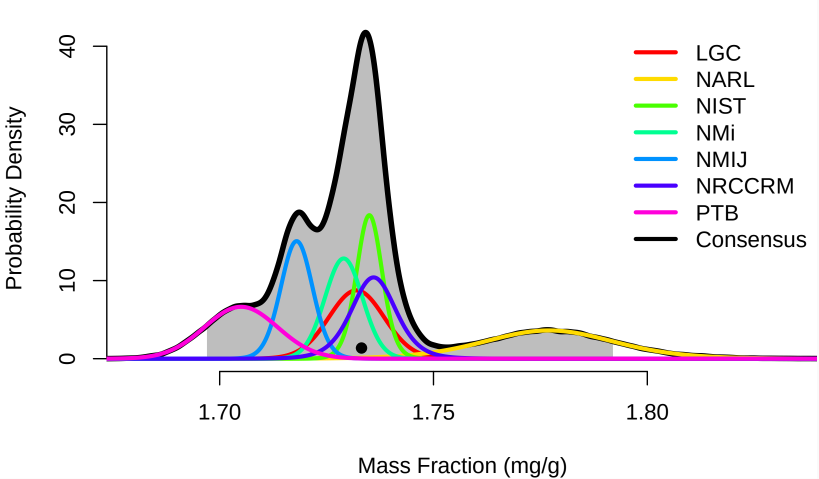

Probability Density
0
5
10
15
1.70
1.75
1.80
Mass Fraction (mg/g)
LGC
NARL
NIST
NMi
NMIJ
NRCCRM
PTB



LGC
NARL
NIST
NMi
NMIJ
NRCCRM
PTB
Probability Density
0
5
10
15
1.70
1.75
1.80
Mass Fraction (mg/g)

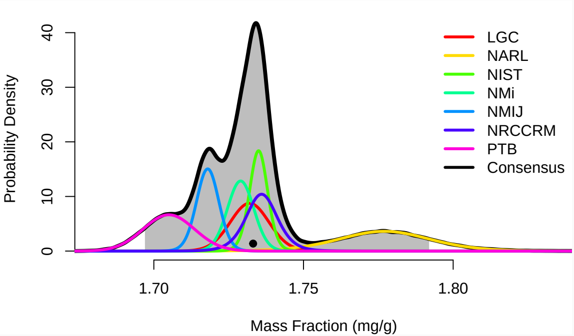


Probability Density
0
10
20
30
40
1.70
1.75
1.80
Mass Fraction (mg/g)
LGC
NARL
NIST
NMi
NMIJ
NRCCRM
PTB
Consensus



Probability Density
0
10
20
30
40
1.70
1.75
1.80
Mass Fraction (mg/g)
LGC
NARL
NIST
NMi
NMIJ
NRCCRM
PTB
Consensus

Linear Pool

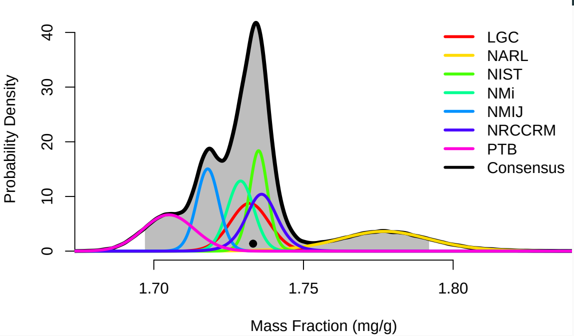


Mixture Model

- $f = \sum_{j=1}^{n} w_j \phi_j$

f	Probability density of measurand
ϕ_j	Probability density for lab j
w_j	Weight of lab j

Linear Pool



Typically, a large sample is drawn from the mixture distribution by repeating the following:

1. Select a laboratory at random
2. Draw a value from the corresponding distribution

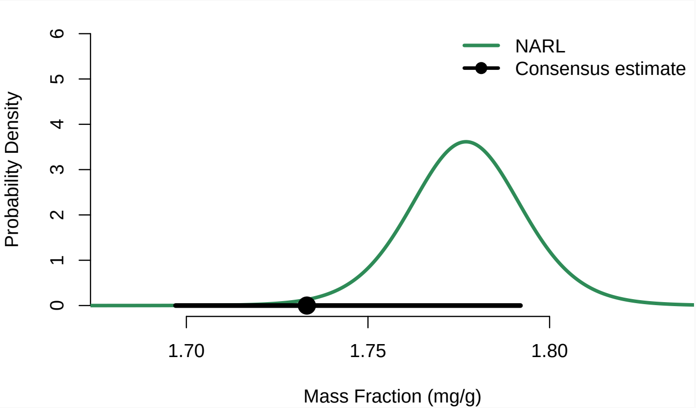
Which method should I use?

Which method should I use?

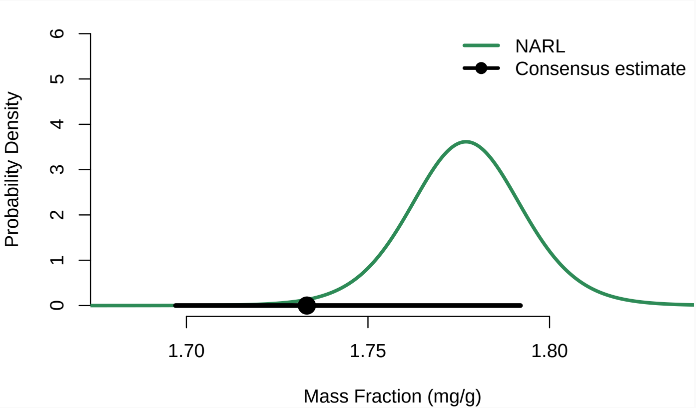
Results for key comparison:

PROCEDURE	CONSENSUS	STD. UNC.	EXP. UNC. $(95\,\%)$
DerSimonian-Laird	1.7294	0.0047	0.0095
Bayesian	1.7291	0.0055	0.0112
Linear Pool	1.7332	0.0222	0.0502

Degrees of Equivalence

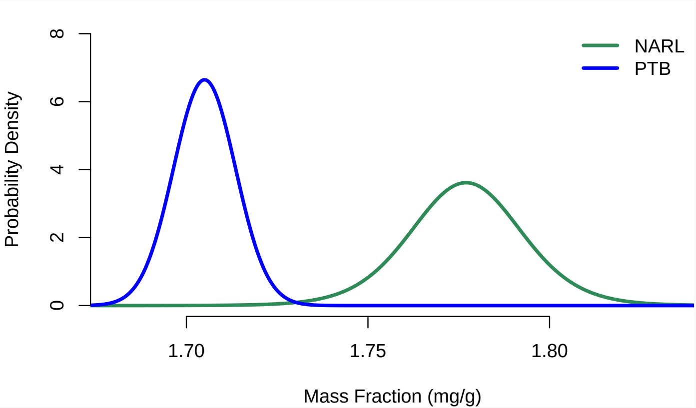


Degrees of Equivalence

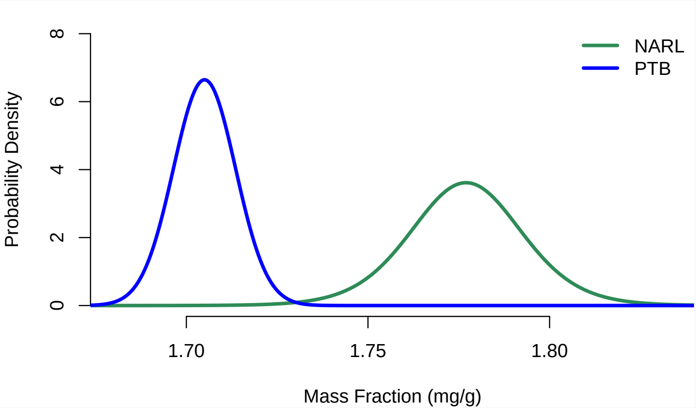


Unilateral DoEs: Identify measurement results that differ significantly from the consensus value

Degrees of Equivalence



Degrees of Equivalence



Bilateral DoEs: Identify measurement results that differ significantly from one another when considered in pairs

Degrees of Equivalence

- Conventional version (as defined by the MRA)
  - Unilateral: $D_j = x_j - \hat{\mu}$
  - Bilateral: $B_{ij} = D_i - D_j$

Degrees of Equivalence

- Conventional version (as defined by the MRA)
  - Unilateral: $D_j = x_j - \hat{\mu}$
  - Bilateral: $B_{ij} = D_i - D_j$
- Uncertainty: Simulate for $k = 1, \ldots, K$

$$D_{j,k} = x_j + e_{j,k} - \hat{\mu},$$

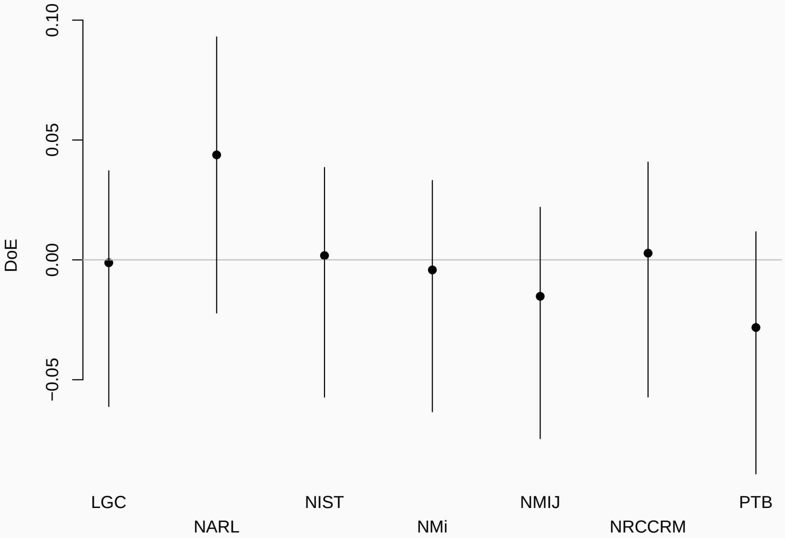
$e_{j,k} \sim$ Student's t with mean 0, variance u_j^2, d.f. ν_j

$$B_{ij,k} = D_{i,k} - D_{j,k}$$

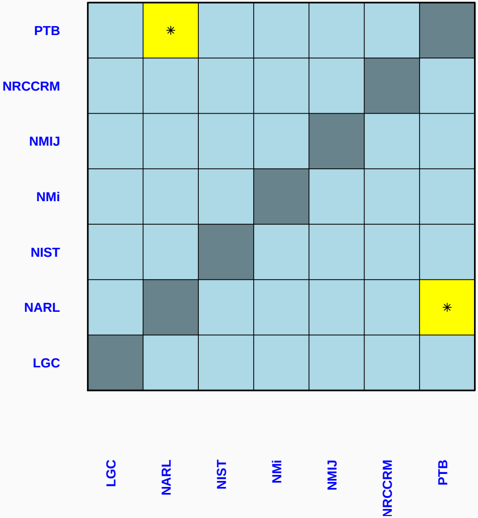
95% coverage interval computed from quantiles of the simulated DoEs.

Unilateral DoEs

DoE estimate and 95% coverage interval



Bilateral DoEs



Yellow squares (with black asterisks in the center) indicate results that differ significantly from 0 at 95% coverage.

Degrees of Equivalence

- Conventional version (as defined by the MRA)
  - Unilateral: $D_j = x_j - \hat{\mu}$
  - Bilateral: $B_{ij} = D_i - D_j$

Degrees of Equivalence

- Conventional version (as defined by the MRA)
  - Unilateral: $D_j = x_j - \hat{\mu}$
  - Bilateral: $B_{ij} = D_i - D_j$
- Leave-one-out version
  - Unilateral: $D_j^* = x_j - \hat{\mu}_{-j}$
  - Bilateral: $B_{ij}^* = D_i^* - D_j^*$

Degrees of Equivalence

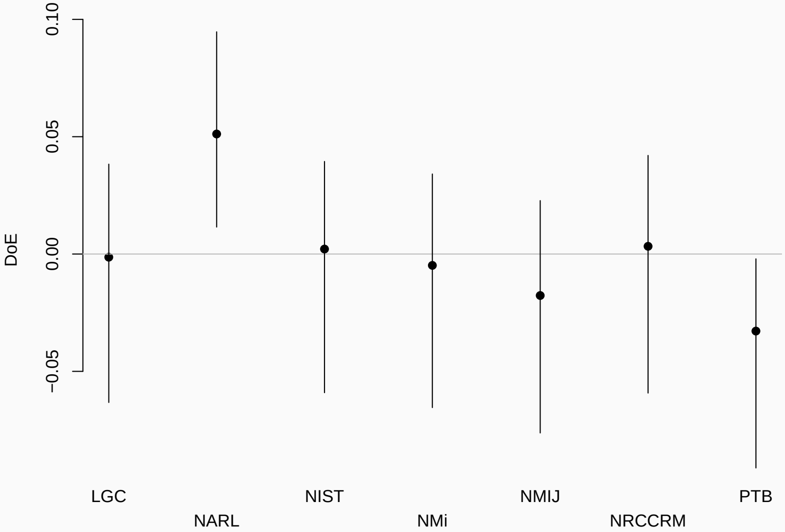
- Leave-one-out version
  - Unilateral: $D_j^* = x_j - \hat{\mu}_{-j}$
  - Bilateral: $B_{ij}^* = D_i^* - D_j^*$

Degrees of Equivalence

- Leave-one-out version
  - Unilateral: $D_j^* = x_j - \hat{\mu}_{-j}$
  - Bilateral: $B_{ij}^* = D_i^* - D_j^*$
- Uncertainty: Simulate for $k = 1, \ldots, K$:
  - $\{\tilde{x}_{-j,k}\}$: Sample from linear pool applied to all but lab j
  - $\{e_{j,k}\}$: Sample from Student's t with mean 0, variance u_j^2, d.f. ν_j
  - $D_{j,k}^* = x_j + e_{j,k} - \tilde{x}_{-j,k}$
  - $B_{ij,k}^* = D_{i,k}^* - D_{j,k}^*$

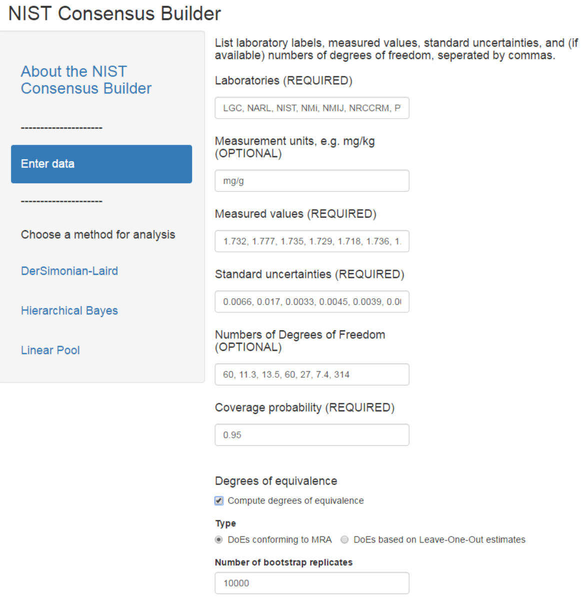
Unilateral DoEs: Leave-one-out version

DoE estimate and 95% coverage interval



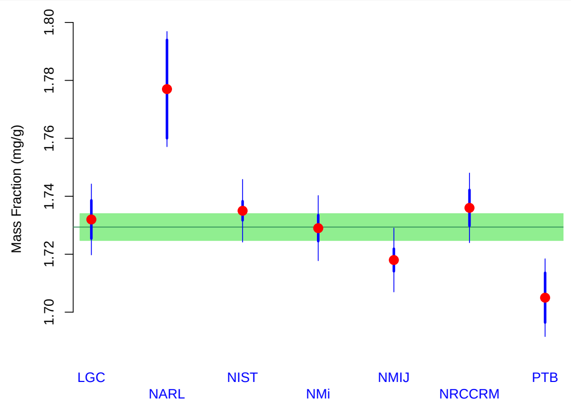
NIST Consensus Builder

NIST Consensus Builder — consensus.nist.gov



Summary

- Goal: Consensus value and uncertainty, degrees of equivalence
- Principles
- Statistical models
- NIST Consensus Builder: consensus.nist.gov



More details: *Consensus building for interlaboratory studies, key comparisons, and meta-analysis*, Koepke et al (2017)