Metrologia **46** (2009) 323–331 doi:10.1088/0026-1394/46/3/021

Weighted means statistics in interlaboratory studies

Andrew L Rukhin

Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-0001, USA

Received 5 February 2009, in final form 6 April 2009
Published 6 May 2009
Online at stacks.iop.org/Met/46/323

Abstract
The usefulness of weighted means statistics as a consensus mean estimator in collaborative studies is discussed. A random effects model designed to combine information from several sources is employed to justify their appeal to metrologists. Some methods of estimating the uncertainties and of constructing confidence intervals are reviewed.

(Some figures in this article are in colour only in the electronic version)

1. Introduction: common mean model for interlaboratory studies

The goal of this paper is to review the use of weighted means statistics in interlaboratory testing. Statistical analysis initiated, for example, when certifying standard reference materials, has the fundamental goal of estimating the overall treatment effect μ (the common effect, the consensus mean or the reference value) and providing a standard error for this estimate. See [1–3] for a detailed discussion of the problem.

Assume there are p laboratories, each measuring the unknown underlying (non-random) value μ common to all laboratories. In the simplest model the measurements x_{ij}, $i = 1, \ldots, p$; $j = 1, \ldots, n_i$, are of the form

$$x_{ij} = \mu + e_{ij}, \tag{1}$$

with independent Gaussian errors $e_{ij} \sim N(0, \kappa_i^2)$. All parameters μ, κ_i^2, $i = 1, \ldots, p$ are unknown, but the main goal is to estimate μ or, more importantly, to provide a confidence interval for μ. The fairly small sample sizes typical in metrology do not always allow asymptotic or non-parametric inference; out of parametric models (1) is the simplest and most widely (albeit not universally) used.

Denote by $x_i = \bar{x}_i = \sum_j x_{ij}/n_i$ the within-lab means and by $s_i^2 = \sum_j (x_{ij} - x_i)^2/[n_i(n_i - 1)]$ (unbiased) estimates of the variances $\sigma_i^2 = \kappa_i^2/n_i$ of x_i. When these variances σ_i^2 are known, the best (in terms of the mean squared error) unbiased estimator of the reference value μ is a weighted means statistic,

$$\tilde{x} = \frac{\sum_i w_i x_i}{\sum_i w_i},$$

with $w_i = w_i^{\mathrm{tr}} = \sigma_i^{-2}$, $i = 1, \ldots, p$. Then the formula for the variance,

$$\mathrm{Var}(\tilde{x}) = E(\tilde{x} - \mu)^2 = \frac{1}{\sum_i w_i^{\mathrm{tr}}}, \tag{2}$$

is well known. These results hold even without the normality assumption if one restricts the class of unbiased estimators to linear unbiased estimators. However, in practice the variances σ_i^2 are *unknown*, so that the 'true' weights w_i^{tr} are also unknown. The usual suggestion [3–5] is to replace σ_i^2 by their estimates s_i^2, i.e. to estimate $\mathrm{Var}(\tilde{x})$ by

$$\left[\sum_i s_i^{-2}\right]^{-1}. \tag{3}$$

Although s_i^2 estimates σ_i^2 unbiasedly, estimate (3) of $\mathrm{Var}(\tilde{x})$ *underestimates* this variance. This fact follows from the inequality,

$$E\left[\sum_i s_i^{-2}\right]^{-1} < \left[\sum_i E s_i^{-2}\right]^{-1},$$

and its implications are known to metrologists who complain that the reciprocal square-root of the sum of the weights becomes too small as the number of participants increases and many labs fall outside the uncertainty interval (see for example [6]). The variation in the s_i^2 themselves, or the uncertainties in the σ_i^2, must be taken into account when estimating the precision of $\tilde{x}$. We will stress this point several times.

The traditional statistical procedure, the maximum likelihood estimator (MLE) of μ, does not have an explicit

form, although it is a weighted means statistic with the weights inversely proportional to the maximum likelihood estimates of σ_i^2. There are numerical algorithms for its evaluation [7–9]. Alternative simpler procedures in our situation include the sample mean $\bar{x}$ and the so-called Graybill–Deal [10] estimator,

$$\tilde{x}_{\mathrm{GD}} = \frac{\sum_i x_i s_i^{-2}}{\sum_i s_i^{-2}}, \tag{4}$$

which merely is the plug-in version of $\tilde{x}$. Estimator (4) is popular among metrologists. In particular, it is used when calculating CODATA recommended values of the fundamental physical constants [11].

An unbiased estimator $\widehat{\mathrm{var}}$ of the variance of $\tilde{x}_{\mathrm{GD}}$ can be expressed in terms of the hypergeometric function [12, pp 194–6]

$$F(1,2;c;z) = \sum_{n=0}^{\infty} \frac{(n+1)!\Gamma(c)}{\Gamma(n+c)} z^n.$$

Namely,

$$\widehat{\mathrm{var}}(\tilde{x}_{\mathrm{GD}}) = \frac{\sum_i \omega_i^{\mathrm{GD}} F(1,2;(n_i+1)/2, 1-\omega_i^{\mathrm{GD}})}{\sum_i 1/s_i^2}.$$

Here $\omega_i^{\mathrm{GD}} = s_i^{-2}/\sum_k s_k^{-2}$ are normalized weights, so that $\tilde{x}_{\mathrm{GD}} = \sum_i \omega_i^{\mathrm{GD}} x_i$.

For $n_i = 3$, $F(1,2;2;1-z) = 1/z$ and

$$\widehat{\mathrm{var}}(\tilde{x}_{\mathrm{GD}}) = \frac{p}{\sum_i 1/s_i^2}.$$

Thus, in this simple situation when all p labs make three measurements, the unbiased estimator is p times larger than estimate (3) of the same parameter. Clearly (3) can dramatically underestimate $\mathrm{var}(\tilde{x}_{\mathrm{GD}})$. A serious drawback of the Graybill–Deal estimator is that small values of s_i^2 lead to unjustifiably large weights. Our simulation results (section 7) confirm that this estimator has serious deficiencies especially for small sample sizes n_i.

Fairweather's estimator [13] of μ is based on the weights $(n_i - 3)/[s_i(n_i - 1)]$,

$$\tilde{x}_F = \frac{\sum_i \dfrac{(n_i-3)}{(n_i-1)s_i} x_i}{\sum_i \dfrac{n_i-3}{(n_i-1)s_i}}. \tag{5}$$

The important feature of this estimator is its relationship to a pivot based on convex combination of t-distributed ratios $(x_i - \mu)/s_i$, which leads to a practical confidence interval determined from a t-approximation with estimated degrees of freedom (see section 5).

Model (1) may not be adequate in situations when the results of different labs do not agree, so that, say, 95% individual confidence intervals for μ based on data from the individual labs do not all overlap. Indeed, it is possible that a lab with the smallest reported uncertainty dominates the data from all other labs. An additional difficulty for (1) arises when one tries to incorporate type B errors of the uncertainty budget. For these reasons more flexible estimators/models are desirable. In the next section we discuss utility of one such model.

2. Random effects model for interlaboratory studies

Assume that the datum x_{ij} in the ith laboratory in addition to the measurement error is affected by a random laboratory effect b_i. More precisely, let

$$x_{ij} = \mu + b_i + \epsilon_{ij}, \tag{6}$$

where, as in (1), $i = 1, \ldots, p$ indexes the laboratories, $j = 1, \ldots, n_i$ represents the sample size (the number of measurements) in laboratory i and μ still is the true mean (reference value). The random variables b_i and ϵ_{ij} are all independent and normal with zero means and variances σ_{B}^2 and τ_i^2; b_i represent the between-laboratory effect (or a hidden error [14]) which is commonly observed in collaborative studies. It is possible that in (6) $b_i \equiv 0$, i.e. $\sigma_{\mathrm{B}}^2 = 0$.

Clearly, (6) leads to the following model for the sample means $x_i = \bar{x}_i = \sum_j x_{ij}/n_i$,

$$x_i = \mu + b_i + e_i. \tag{7}$$

Here $b_i \sim N(0, \sigma_{\mathrm{B}}^2)$ and $e_i \sim N(0, \sigma_i^2)$ are mutually independent.

Cochran [15] introduced this model in 1937 (see [16] for a review). He studied the MLE which, as for (1), does not admit an explicit form. He reports results of an early numerical efficiency study in which the sample mean, $\bar{x}$, the Graybill–Deal estimator, $\tilde{x}_{\mathrm{GD}}$, and the MLE were compared when $\sigma_{\mathrm{B}}^2 = 0$. Cochran writes 'when p is as low as 6, MLE is satisfactory, but tedious' to evaluate, and 'there is little to choose between $\tilde{x}_{\mathrm{GD}}$ and $\bar{x}$, but occasionally $\tilde{x}_{\mathrm{GD}}$ wins handsomely'; $\tilde{x}_{\mathrm{GD}}$ 'may be recommended' when $p \geqslant 15$. The sample mean is better than $\tilde{x}_{\mathrm{GD}}$ when p is fairly small and σ_i do not vary much. The results of a similar study for some positive values of σ_{B}^2 are given in [17].

Because of the rather inconclusive nature of such studies caused by the large number of parameters and complicated form of the likelihood equations, simpler procedures are desirable in practice. Estimators of the common mean via moment-type equations are reviewed in the next section.

Before that we note that model (6) may not help to understand possible systematic influences on the measurement results of one or several labs [18]. Still, a common distribution of hidden errors applicable to all labs clarifies further uncertainty analysis. Both models (1) and (6) have been criticized by metrologists as they assume potentially unresolved differences through an infinite population of laboratories/institutes while in many interlaboratory studies (especially in the so-called Key Comparisons) there is only a limited number of qualified participants [19, 20]. However, taking into account the exact nature of all laboratory measurement techniques needed in the formulation of a finite population sampling is difficult, if not impossible; the alternative finite population sampling models lead to less tractable mixture type distributions, while the relationship

of (6) and error estimation theory in linear models benefits evaluation of the ensuing uncertainties.

3. Estimating equations and weighted means statistics

In model (6) the within-labs variances σ_i^2 can be estimated by the available estimates s_i^2 (type A uncertainty), but the problem of estimating the between-study component of variance σ_B^2 remains. Here several estimators of $\mathrm{Var}(\tilde{x})$ for a class of weighted means statistics $\tilde{x}$ of the form

$$\tilde{x} = \frac{\sum_i \dfrac{x_i}{y + s_i^2}}{\sum_i \dfrac{1}{y + s_i^2}} \tag{8}$$

are suggested. Here y is supposed to estimate the unknown variance σ_B^2. Thus, we restrict our attention to the weights of the form

$$w_i = \frac{1}{y + s_i^2}. \tag{9}$$

Because of positive y, (9) is much less sensitive than the Graybill–Deal weights to small values of s_i^2. Indeed the presence of y makes it impossible for one laboratory to dominate all others unless all labs produce similar results (in which case σ_B^2 is estimated by zero.) The limiting case $y = \infty$ corresponds to the arithmetic (sample) mean with equal weights.

An estimating equation was suggested to get an estimator (8) of μ. If the weights w_i are arbitrary,

$$E\sum_i w_i(x_i - \tilde{x})^2 = \sum_i (\sigma_B^2 + \sigma_i^2)w_i - \frac{\sum_i (\sigma_B^2 + \sigma_i^2)w_i^2}{\sum_i w_i}$$

$$= \sigma_B^2\left[\sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}\right] + \sum_i \sigma_i^2 w_i - \frac{\sum_i \sigma_i^2 w_i^2}{\sum_i w_i} \tag{10}$$

[21, 22]. In particular, when $w_i = 1/\sigma_i^2$,

$$E\sum_i \frac{(x_i - \tilde{x})^2}{\sigma_i^2} = p - 1 + \sigma_B^2\left[\sum_i \frac{1}{\sigma_i^2} - \frac{\sum_i \dfrac{1}{\sigma_i^4}}{\sum_i \dfrac{1}{\sigma_i^2}}\right]. \tag{11}$$

By employing the idea behind the method of moments, DerSimonian and Laird [23] made use of identity (11) as an estimating equation for μ and σ_B^2 in the following way. Determine a non-negative $y = y_{DL}$ from the formula

$$\sum_i \frac{(x_i - \tilde{x}_{GD})^2}{s_i^2} = p - 1 + y\left[\sum_i s_i^{-2} - \frac{\sum_i s_i^{-4}}{\sum_i s_i^{-2}}\right],$$

i.e. with the Graybill–Deal estimator $\tilde{x}_{GD}$ in (4),

$$y_{DL} = \max\left[0, \frac{\sum_i s_i^{-2}(x_i - \tilde{x}_{GD})^2 - p + 1}{\sum_i s_i^{-2} - \sum_i s_i^{-4}\left[\sum_i s_i^{-2}\right]^{-1}}\right].$$

Thus, the statistic $\tilde{x}_{GD}$ and the weights $w_i = s_i^{-2}$, corresponding to $\sigma_B^2 = 0$, are used to estimate $E\sum_i (x_i - \tilde{x})^2/\sigma_i^2$, which then serves to find the true σ_B^2 via (11).

The resulting estimator,

$$\tilde{x}_{DL} = \frac{\sum_i \dfrac{x_i}{y_{DL} + s_i^2}}{\sum_i \dfrac{1}{y_{DL} + s_i^2}}, \tag{12}$$

became immensely popular especially in biostatistics. DerSimonian and Laird, motivated by (2), also gave an approximate formula for the estimate of the variance of $\hat{x}_{DL}$,

$$\widehat{\mathrm{Var}}(\tilde{x}_{DL}) = \frac{1}{\sum_i (y_{DL} + s_i^2)^{-1}},$$

which is similar to (3).

The Mandel–Paule algorithm [24, 25] uses weights of the form (9) as well. However now $y = y_{MP}$, which is designed to approximate σ_B^2, is found from the moment-type estimating equation,

$$F(y_{MP}) = p - 1, \tag{13}$$

where with $\tilde{x}$ defined by (8),

$$F(y) = \sum_i \frac{(x_i - \tilde{x})^2}{y + s_i^2}$$

is a convex monotonically decreasing function of $y \geqslant 0$. Motivation for (13) comes from the formula

$$E\sum_i w_i^{tr}(x_i - \tilde{x})^2 = p - 1,$$

which follows from (11) when the weights w_i are optimal, i.e. when they coincide with w_i^{tr}. The explicit solution of (13) for $p \geqslant 3$ does not exist; in practice a number of iterations is needed to get it with desired accuracy. The following approximation is easily computable:

$$\hat{x}_{MPA} = \frac{\sum_i \dfrac{x_i}{y_{MPA} + s_i^2}}{\sum_i \dfrac{1}{y_{MPA} + s_i^2}}, \tag{14}$$

where

$$y_{MPA} = \begin{cases} y_{DL} + \dfrac{F(y_{DL})}{|F'(y_{DL})|} & \text{if } \dfrac{2F(y_{DL})F''(y_{DL})}{[F'(y_{DL})]^2} \geqslant 1 \\ y_{DL} + \dfrac{|F'(y_{DL})|}{F''(y_{DL})} - \sqrt{\left[\dfrac{F'(y_{DL})}{F''(y_{DL})}\right]^2 - \dfrac{2F(y_{DL})}{F''(y_{DL})}} & \\ \text{otherwise.} & \end{cases}$$

The formula for the derivative of the weighted sum of squares [26, p 323] shows that for example,

$$F'(y_{DL}) = -\sum_i \frac{(x_i - \hat{x}_{DL})^2}{(y_{DL} + s_i^2)^2},$$

and this solution is the one-step application of the Newton method for the initial value $y = y_{DL}$. Notice that $(p - 1)^{-1}\sum_i (x_i - \hat{x}_{MP})^2/(y_{MP} + s_i^2)$ is the square of the so-called Birge ratio which is commonly used in metrology for testing goodness-of-fit. Thus, the Mandel–Paule procedure seeks the

weights under which the squared Birge ratio equals its expected value. Schiller and Eberhardt [27] write about the Mandel–Paule method: '... seems to be about the best scheme available'.

The modified Mandel–Paule procedure with $y = y_{\mathrm{MMP}}$ is defined by replacing $p-1$ in the right-hand side of (13) by p, i.e.

$$\sum_i \frac{(x_i - \tilde{x}_{\mathrm{MMP}})^2}{y_{\mathrm{MMP}} + s_i^2} = p. \tag{15}$$

As was shown in [28], this procedure is characterized by the following fact: the MLE $\hat{\sigma}_{\mathrm{B}}^2$ of σ_{B}^2 coincides with y_{MMP}, if in the reparametrized version of the likelihood equation the weights w_i admit representation (9).

Thus, the modified Mandel–Paule estimator can be interpreted as a procedure which uses the weights of the form $1/(y + s_i^2)$ (instead of solutions of the likelihood equation that are difficult to find) and still maintains the same estimate of σ_{B}^2 as the maximum likelihood. A similar interpretation holds for the original Mandel–Paule rule and the restricted likelihood function. For this reason both Mandel–Paule rules are natural approximations of their maximum likelihood counterparts. The multivariate extension of these two methods is also available [29].

4. Behaviour of weighted means: large number of labs

Here we look at the behaviour of the class of statistics that includes the DerSimonian–Laird procedure and the Mandel–Paule rule assuming (perhaps rather unrealistically) that the number p of different laboratories is large. The class is composed of general weighted means statistics $\tilde{x}$ of the form (8) with w_i given by (9). The value of y is determined from an estimating equation such as (12) or (13). Under the assumptions detailed below, this quantity converges with probability one to a constant obtained from the limiting form of the estimating equations.

We regard the variances, σ_i^2 as i.i.d. (independent identically distributed) realizations of a random variable with some fixed but otherwise arbitrary distribution function G. Although in practice the elicitation of G from practitioners is difficult, this approach is useful since approximate variance estimation for the statistics (8) becomes possible.

Let the observable i.i.d. random variables x_i, s_i^2, $i = 1, 2, \ldots, p$, be realizations of the random vector (X, S^2) such that X and S^2 are conditionally (for given σ) independent with the conditional distribution of X being $N(\mu, \sigma_{\mathrm{B}}^2 + \sigma^2)$ for some unknown σ_{B}^2. For simplicity, we take both μ and σ_{B}^2 to be fixed. The conditional distribution of S^2 is supposed to be of the form $\sigma^2 W$ with a random variable W, $EW = 1$, which is independent of σ^2. In the typical Gaussian case W has the distribution of χ_ν^2/ν with ν being the typical degrees of freedom or a mixture of such distributions. A similar model has been used when $\sigma_{\mathrm{B}}^2 = 0$ [30].

Thus, $E(X|\sigma) = \mu$ and $E([X-\mu]^2|\sigma) = \sigma_{\mathrm{B}}^2 + \sigma^2$. The law of large numbers shows that for a fixed y,

$$\frac{1}{p}\sum_i \frac{1}{y + s_i^2} \to E\frac{1}{y + S^2}$$

and for a fixed non-negative y,

$$\tilde{x} = \frac{\sum_i \dfrac{x_i}{y + s_i^2}}{\sum_i \dfrac{1}{y + s_i^2}} \to \frac{E\dfrac{X}{y + S^2}}{E\dfrac{1}{y + S^2}} = \frac{E\left[E(X|\sigma)E\left(\dfrac{1}{y + S^2}|\sigma\right)\right]}{E\left[\left(E\dfrac{1}{y + S^2}|\sigma\right)\right]} = \mu.$$

Thus, under our assumptions, $\tilde{x}$ is a consistent estimator of μ, and, according to the Central Limit Theorem, $p^{-1/2}\sum_i w_i(x_i - \mu) = p^{-1/2}(\tilde{x} - \mu)\sum_i w_i$ has an approximately normal distribution with zero mean and with the variance $E(X-\mu)^2(y + S^2)^{-2}$. Therefore, $p^{1/2}(\tilde{x} - \mu)$ is asymptotically normally distributed with zero mean and the variance

$$S(y) = \frac{E\dfrac{(X-\mu)^2}{(y + S^2)^2}}{\left(E\dfrac{1}{y + S^2}\right)^2} = \frac{E\dfrac{\sigma_{\mathrm{B}}^2 + \sigma^2}{(y + \sigma^2 W)^2}}{\left(E\dfrac{1}{y + \sigma^2 W}\right)^2} \geqslant \frac{1}{E\dfrac{1}{\sigma_{\mathrm{B}}^2 + \sigma^2}}. \tag{16}$$

For moderate p, $p \leqslant 15$, when y is small, this normal approximation may be inadequate. Given the distributions of W and σ, the asymptotically optimal value of $y = y_{\mathrm{opt}}$ can be found as the minimizer of $S(y)$. Observe that if $W \equiv 1$, $y_{\mathrm{opt}} = \sigma_{\mathrm{B}}^2$, in which case the lower bound in (16) is attained.

When $\sigma^2 \equiv \sigma_0^2$, $S(y)$ monotonically decreases to the value $\sigma_{\mathrm{B}}^2 + \sigma_0^2$, so that in this case $y_{\mathrm{opt}} = \infty$, and $\bar{x}$ is asymptotically optimal. In general, for $y \to \infty$,

$$\frac{S'(y)}{2E(\sigma_{\mathrm{B}}^2 + \sigma^2)} \sim \frac{\mathrm{Var}(\sigma^2)}{y^2} > 0,$$

so that $S(y)$ increases for large y, and then $y_{\mathrm{opt}} < \infty$.

Also, provided that $E\sigma^{-4} < \infty$, $S'(0) < 0$, unless $W \equiv 1$. Therefore, in this setting, the Graybill–Deal estimator with $y = 0$ cannot be optimal for non-degenerate distributions of W. For a fixed positive y, the variance of $\tilde{x}$ can be estimated via a consistent estimate of $S(y)$, e.g. by

$$\delta_0 = \frac{p}{p-1}\sum_i \frac{(x_i - \tilde{x})^2}{(y + s_i^2)^2}\left[\sum_i \frac{1}{y + s_i^2}\right]^{-2} \tag{17}$$

or by $\sum_i \frac{(x_i-\tilde{x})^2}{(y+s_i^2)^2}\left[\sum_i \frac{1}{y+s_i^2}\right]^{-2}$ [28]. The factor $p(p-1)^{-1}$ in (17) is motivated by the fact that (17) corresponds to an unbiased estimator, $\sum_i (x_i - \tilde{x})^2/[p(p-1)]$, of the variance of the sample mean $\bar{x}$ when all σ_i^2 are equal.

5. Confidence intervals based on the weighted means

If z_α denotes the critical point of the standard normal distribution, for large p the interval,

$$\tilde{x} \pm z_{\alpha/2}\frac{\sqrt{p\sum_i \dfrac{(x_i - \tilde{x})^2}{(y + s_i^2)^2}}}{\sqrt{p-1}\sum_i \dfrac{1}{y + s_i^2}}, \tag{18}$$

is an approximate $(1-\alpha)100\%$-confidence interval for μ on the basis of the weighted means statistics $\tilde{x}$. In practice it is reasonable to replace the critical point $z_{\alpha/2}$ by that of the t-distribution, $t_{\alpha/2}(p-1)$.

We stress again that for larger p, the variance of the Mandel–Paule rule $\tilde{x}$ is better estimated by (17) with $y = y_{\mathrm{MP}}$ determined by (13), rather than by (3) as suggested by Mandel [1, p 72]. However, Mandel writes: $(y + s_i^2)^{-1}$ 'are actually only sample estimates of the true weights resulting in perhaps considerable uncertainty in' $\delta_1 = [\sum_i (y + s_i^2)^{-1}]^{-1}$. In the setting of section 4 these estimators cannot give a good estimate of the variance of the weighted means statistic with weights (9), as this would suggest that the minimal value of the variance is attained at $y = 0$. An alternative estimator of the variance of $\tilde{x}$ can be obtained for any p from the following procedure suggested in the context of general linear models [31].

Let $\omega_i = w_i/(\sum_k w_k)$, $\sum_i \omega_i = 1$, be fixed normalized weights, which determine the weighted means statistic, $\tilde{x} = \sum_i \omega_i x_i$, with the variance, $\mathrm{Var}(\tilde{x}) = \sum_i \omega_i^2 \mathrm{Var}(x_i)$. For the (unbiased) weighted means statistic $\tilde{x}$,

$$E(x_k - \tilde{x})^2 = (1 - 2\omega_k)\mathrm{Var}(x_k) + \sum_i \omega_i^2 \mathrm{Var}(x_i).$$

When $\omega_i = \omega_i^{\mathrm{tr}} = w_i^{\mathrm{tr}}(\sum_k w_k^{\mathrm{tr}})^{-1}$, $\tilde{x}$ is the optimal least squares estimator, and the second term in the right-hand side simplifies to $[\sum_i \mathrm{Var}(x_i)^{-1}]^{-1} = \omega_k^{\mathrm{tr}}\mathrm{Var}(x_k)$. By substituting this expression, one obtains

$$E(x_k - \tilde{x})^2 = (1 - \omega_k^{\mathrm{tr}})\mathrm{Var}(x_k).$$

Horn *et al* [31, p 382] argue that by continuity, if the weights are close to ω_k^{tr}, this is an approximate identity. Thus, one derives an *almost unbiased* estimator of $\mathrm{Var}(x_k)$ as $(x_k - \tilde{x})^2/(1-\omega_k)$, and the corresponding estimate of the variance, $\mathrm{Var}(\tilde{x})$, is

$$\widehat{\mathrm{Var}}(\tilde{x}) = \sum_i \frac{\omega_i^2 (x_i - \tilde{x})^2}{1 - \omega_i}.$$

This statistic gives an estimate of the variance of any weighted means statistic for weights (9) when s_i^2 are fixed. The method leads to the following estimate δ_2 of $\mathrm{Var}(\tilde{x})$,

$$\delta_2 = \frac{\sum_i \dfrac{(x_i - \tilde{x})^2}{(y + s_i^2)^2} \left[\sum_{k:k\neq i} \dfrac{1}{y + s_k^2}\right]^{-1}}{\sum_i \dfrac{1}{y + s_i^2}}. \qquad (19)$$

with the plug-in weights $\omega_i = (y + s_i^2)^{-1}/\sum_k (y + s_k^2)^{-1}$, $i = 1, \ldots, p$. Simulations show that (19) gives good confidence intervals of the form $\tilde{x} \pm t_{\alpha/2}(p-1)\sqrt{\delta_2}$. For the Mandel–Paule rule or the DerSimonian–Laird procedure they outperform the intervals $\tilde{x} \pm t_{\alpha/2}(p-1)\sqrt{\delta_1}$.

Estimator (19) alleviates the problem mentioned in section 1 for the Graybill–Deal estimator when one laboratory reports a very small uncertainty. Then not only this laboratory estimate becomes $\tilde{x}_{\mathrm{GD}}$, but also its uncertainty takes over as the estimate of the variance of this statistic.

Indeed if, say, $s_1^2 \ll s_i^2$, $i = 2, \ldots, p$, then (3) practically coincides with s_1^2. However according to (19),

$$\widehat{\mathrm{Var}}(\tilde{x}_{\mathrm{GD}}) \simeq s_1^2 \left(\sum_{i=2}^{p} \frac{x_i - x_1}{s_i^2}\right)^2 \left[\sum_{i=2}^{p} \frac{1}{s_i^2}\right]^{-1}.$$

The data-dependent factor on the right-hand side of this formula typically prevents $\widehat{\mathrm{Var}}(\tilde{x}_{\mathrm{GD}})$ from getting very close to s_1^2.

We conclude this section with another procedure based on a quadratic estimator $\sum_i q_i(x_i - \tilde{x})^2$ of the variance of a weighted means statistic $\tilde{x}$. A natural confidence interval for μ based on $\tilde{x}$ is $\tilde{x} \pm t\sqrt{\sum_i q_i(x_i - \tilde{x})^2}$, and the question is an appropriate choice for t. When t is large,

$$t^{p-1} \sup_{\sigma_1^2,\ldots,\sigma_p^2} P\left(\left|\frac{\tilde{x} - \mu}{\sqrt{\sum_i q_i (x_i - \tilde{x}y)^2}}\right| > t\right)$$

$$\simeq t^{p-1} P\left(|T_{p-1}| > t\sqrt{(p-1)\left(\gamma p^p \prod_i q_i\right)^{1/(p-1)}}\right)$$

$$= \frac{\Gamma(p/2)}{\sqrt{\pi}\,\Gamma((p+1)/2)(p-1)^{p-1} p^p \gamma \prod_i q_i},$$

where $\gamma = \sum_i \omega_i^2/q_i \geqslant 1/q$, $q = \sum q_i$, and T_{p-1} denotes a t-random variable with $p-1$ degrees of freedom [32]. In other words, the smallest coverage probability of the $(1-\alpha)100\%$-confidence interval, $\tilde{x} \pm t\sqrt{\sum_i q_i(x_i - \tilde{x})^2}$, when α is small, is attained for a t-distribution with $p-1$ degrees of freedom. The 'least-favourable' variances σ_i^2 are all equal.

The shortest interval obtains when $\omega_i = q_i/q$, and this interval,

$$\tilde{x} \pm \frac{t_{\alpha/2}(p-1)\sqrt{\sum_i \omega_i (x_i - \tilde{x})^2}}{\sqrt{(p-1)\left[p^p \prod_i \omega_i\right]^{1/(p-1)}}}, \qquad (20)$$

can be recommended in practice especially when $\min_i n_i < 5$. Under model (1) the confidence interval based on the Fairweather procedure (5) has the average width smaller than (20), but this dominance disappears in a more general situation of (6). Notice that the conservative interval (20) is wider than the interval defined by the so-called external consistency estimator of the variance, $(p-1)^{-1}\sum_i \omega_i (x_i - \tilde{x})^2$ [33].

6. Type B uncertainty and Bayes estimators

Model (6) can be adjusted to incorporate type B uncertainty. More precisely, assume now that the data x_{ij} have the form

$$x_{ij} = \mu + \lambda_i + b_i + \epsilon_{ij}. \qquad (21)$$

The random variables b_i, ϵ_{ij} are still assumed to be mutually independent and normal with zero means and variances σ_{B}^2 and τ_i^2, respectively. The component λ_i represents the type B uncertainty assessed by the laboratory i as composed of a systematic bias component, δ_i, and a variance component, β_i^2.

Define λ_i in a hierarchical way, $\lambda_i \mid \xi_i \sim N(\xi_i, \beta_i^2)$, with the expected bias component ξ_i for lab i being normal

$N(0, \varphi_i^2)$, so that $E\lambda_i = 0$, $\mathrm{Var}(\lambda_i) = \varphi_i^2 + \beta_i^2 = \sigma_{\mathrm{B}_i}^2$. Assume that the reported combined standard type B uncertainty provides an estimate of the variance, $\sigma_{\mathrm{B}_i}^2$, whereas the individual estimates of φ_i^2 and β_i^2 are not available. Then type B uncertainty, $\sigma_{\mathrm{B}_i}^2$, becomes merely a variance component, which can be added to s_i^2 in all formulas in sections 3–5. Of course if $E\lambda_i \neq 0$, then all weighted means statistics become biased, and μ itself cannot be estimated. Thus, we assume that all recognized systematic errors (biases) have been corrected for as recommended [34]. Rukhin and Sedransk [35] discuss metrological implications of models (1), (6) and (21), which also can be interpreted using the Bayesian paradigm.

The (generalized) Bayes estimator of μ under the squared error loss is approximately a weighted means statistic when the prior distribution has the following structure. Take a 'non-informative' improper prior, i.e. a constant density for μ, and some prior density π of the remaining parameters $\sigma_1, \ldots, \sigma_p$, σ_{B}. The formula for this estimator is

$$\tilde{x}_\pi(x_1, \ldots, x_p, s_1, \ldots, s_p) = \frac{\sum_i x_i \gamma_i}{\sum_i \gamma_i},$$

where $\gamma_i = E(\sigma_{\mathrm{B}}^2(\sigma_i^2 + \sigma_{\mathrm{B}}^2)^{-1} | x_1, \ldots, x_p, s_1, \ldots, s_p)$.

Choice of the prior density π such that explicit calculation of γs can be performed does not seem to be feasible. However, if the density π is fairly flat and the likelihood integrated over μ is peaked, one can approximate δ_π by the weighted means statistic $(\sum_i x_i \tilde{\gamma}_i)/(\sum_i \tilde{\gamma}_i)$. Here $\tilde{\gamma}_i$ minimize the function,

$$\left(\sum n_i - 1\right) \log\left[\sum_i \gamma_i (x_i - \tilde{x})^2 + \sum_i \frac{\gamma_i (n_i - 1) s_i^2}{1 - \gamma_i}\right] + \log\left(\sum_i \gamma_i\right) + \sum_i (n_i - 1) \log(1 - \gamma_i) - \sum_i n_i \log \gamma_i.$$

These weights coincide with with the restricted maximum likelihood solution mentioned in section 3.

7. Simulation results

The results of a Monte Carlo simulation study for $p = 5$, 12, 25 and randomly chosen sample sizes n_i with the uniform distribution over integers from 4 to 12 are reported here as a function of $\sigma_{\mathrm{B}}^2 = 0, 1, \ldots, 10$. The error variances σ_i^2 were taken to have a lognormal distribution, so that $E\sigma_i^2 = 1$.

The MLE and the restricted maximum likelihood (REML) estimator were computed via their *R*-language implementation (through the *lme* function from *nlme* library). The *intervals* function with fixed effects provides approximate confidence intervals for μ. The employed formula for the variance of these two estimators is based on the observed Fisher information, which is $1/\sum \hat{\sigma}_i^{-2}$ [7], i.e. it is similar to δ_1 with the MLEs $\hat{\sigma}_i^2$ replacing s_i^2. The simulation results indicate that this (essentially asymptotic in n_i) formula can seriously underestimate the true variance of the MLE (when $n_i \simeq 8$), i.e. the length Δ of these intervals might be too short. For multimodal or flat likelihood functions convergence of the algorithm is problematic. Figure 1 depicts a clearly non-normal q–q plot of pivotal quantity $(\mathrm{MLE} - \mu)/\Delta$ when $p = 5$ and $\sigma_{\mathrm{B}}^2 = 0$ with 50 000 runs.

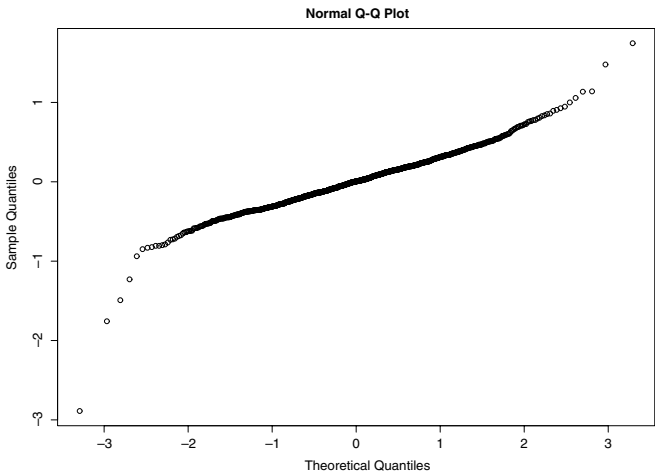

Figure 1. The q–q plot of the pivots for the maximum likelihood estimator when $p = 5$, $\sigma_{\mathrm{B}}^2 = 0$.

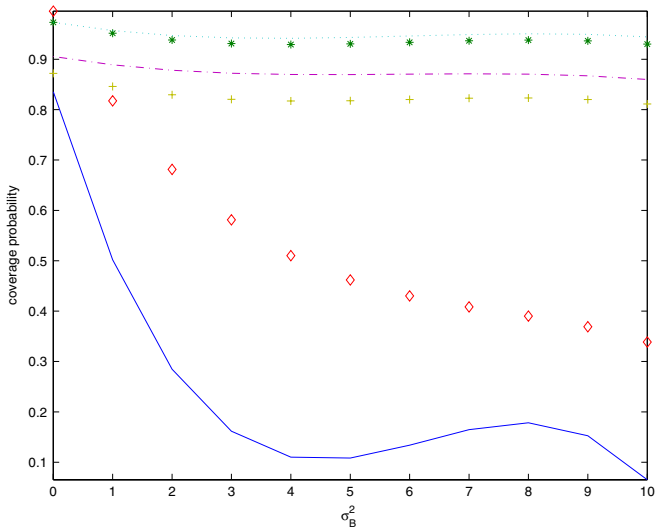

Figure 2. Plot of actual coverage probabilities for confidence intervals with the nominal 95% level based on MLE (line marked by +), REML (dashed–dotted line), x_{GD} (continuous line), x_{DL} (line marked by $*$), x_{MP} (dotted line), x_{F} (line marked by $\Diamond$), when $p = 5$.

The coverage probability of the intervals based on MLE and REML when $p = 5$ did not exceed 91%, staying about 82% (MLE) and about 87% (REML) for most σ_{B}^2 values. These two intervals exhibit better performance in the balanced case $n_i \equiv n$, but then both the sample mean and the Fairweather procedure outperform them.

Figure 2 displays the coverage probability of these intervals with a nominal confidence coefficient of 95% when $p = 5$ and the variance estimator is δ_1. Both the DerSimonian–Laird estimator (12) and the Mandel–Paule procedure with $y = y_{\mathrm{MPA}}$ sustain this confidence level very well. The Graybill–Deal estimator (4) cannot be recommended especially with estimate (3) as its coverage probability drops almost to zero for large σ_{B}^2.

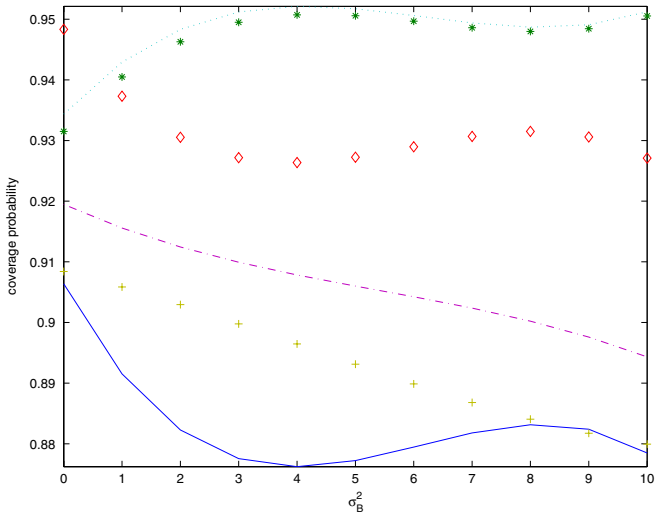

Figure 3. Plot of coverage probabilities of the confidence intervals based on δ_2 when $p = 12$ (designations of lines are the same as in figure 2).

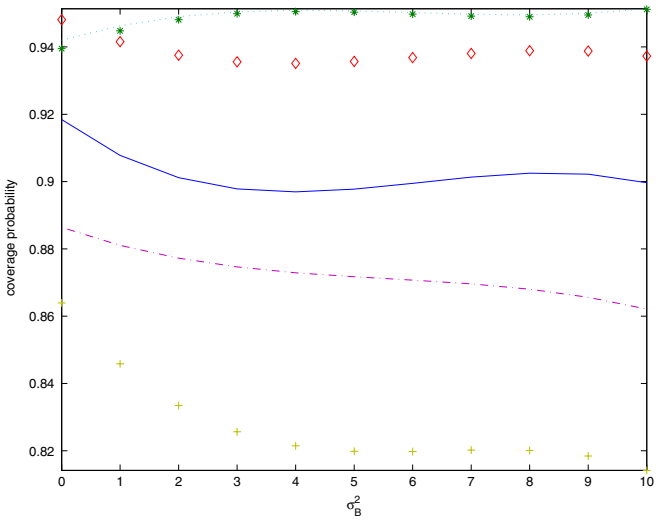

Figure 4. Plot of coverage probabilities of the confidence intervals based on δ_2 when $p = 25$ (designations of lines are the same as in figure 2).

The Fairweather estimator, $\tilde{x}_F$, is reasonable when δ_2 is used, especially for small σ_B^2, but poor with δ_0 and δ_1. Figures 3 and 4 show the coverage probabilities of the same intervals with a nominal confidence coefficient of 95% when $p = 12$ and 25 for the variance estimator δ_2. The average half-widths (standard errors) of these intervals are increasing as σ_B^2 increases, but in the case of likelihood estimators not fast enough to compensate for the loss in stated confidence.

8. Examples

8.1. Determination of Newton's gravitational constant

In the first example we compare two studies (1998 and 2002) involving the Newtonian gravitational constant reported in [11, 36]. The data are given in tables 1 and 2. In table 1 the studies have the following numbering: CODATA-86 = 1, PTB-95 = 2, LANL-97 = 3, TR&D-98 = 4, JILA-98 = 5, HUST-99 = 6, MSL-99 = 7, BIPM-99 = 8, UZur-99 = 9, UWup-99 = 10.

In the 1998 study the outlying result of laboratory 2 influences the $\tilde{x}_{GD}$ to take the value 6.6818, while $\tilde{x}_{MP} = 6.6795$ and $\tilde{x}_{DL} = 6.6796$. The approximate 95%-confidence intervals based on estimates (19) are (6.6695, 6.6897) for $\tilde{x}_{DL}$ and (6.6690, 6.6899) for $\tilde{x}_{MP} = 6.6795$.

The interval based on $\tilde{x}_{GD}$ is quite narrow: (6.6812, 6.6823). The problem with this interval becomes clear after inspecting the 2002 data given in table 2 [11], where LANL-97 = 1, TR&D-98 = 2, HUST-99 = 3, UWash-00 = 4, BIPM-01 = 5, UWup-02 = 6, UZur-02 = 7, MSL-03 = 8.

The 2002 value $\tilde{x}_{GD} = 6.6742$ was not covered by the interval above. In hindsight the DerSimonian–Laird procedure (as well as the Mandel–Paule rule) is more robust to the outlying result, and the 2002 value is in agreement with the advocated confidence intervals on the basis of 1998 data. Neither the maximum likelihood estimator nor the Fairweather estimator are available, because in this example (as in many others) the sample sizes n_i were not specified.

8.2. Gas concentration estimation

Next is an example from analytical chemistry data from gas metrology international comparisons [37] which gave the average concentrations and uncertainties in $\mu\text{mol}\,\text{mol}^{-1}$ as shown in table 3.

While the consensus values evaluated according to different methods were rather close

$$\bar{x} = 10.0749, \qquad \tilde{x}_F = 10.0262,$$

$$\tilde{x}_{GD} = \tilde{x}_{DL} = \tilde{x}_{MP} = \tilde{x}_{MMP} = 10.0225,$$

($y_{MP} = y_{DL} = 0$), their estimated expanded uncertainties were felt by specialists to be too small:

$$\delta_0 = \delta_1 = \left[\sum_i s_i^{-2}\right]^{-1} \approx 0.001.$$

In this situation, interval (18) may not be appropriate, but the conservative interval (20) gives a very sensible answer: 10.0262 ± 0.0907 (the Mandel–Paule estimator), 10.0225 ± 0.0919 (the DerSimonian–Laird estimator). Again the Fairweather interval is not available, because in this example $n_{min} = 3$.

9. Summary and conclusions

The weighted means estimators of the common mean have many desirable statistical features: they are unbiased and consistent, they have properties of asymptotic efficiency and can be easily evaluated. These estimators lead to t-distribution based confidence intervals (20), admit a Bayesian interpretation, and allow adjustments to incorporate type B uncertainty.

However, the mentioned properties are in full play only if the variance of such a procedure is carefully estimated. The maximum likelihood intervals produced in *R* language can

Table 1. 1998 data on the Newtonian gravitational constant, $p = 10$, x_i are measured in $m^3\,kg^{-1}\,s^{-2} \times 10^{-11}$, s_i in $m^3\,kg^{-1}\,s^{-2} \times 10^{-13}$.

	i									
	1	2	3	4	5	6	7	8	9	10
x_i	6.673	6.715	6.674	6.673	6.687	6.670	6.674	6.683	6.675	6.673
s_i	0.085	0.056	0.07	0.05	0.94	0.07	0.07	1.1	0.15	0.29

Table 2. 2002 data on the Newtonian gravitational constant, $p = 8$, x_i are measured in $m^3\,kg^{-1}\,s^{-2} \times 10^{-11}$, s_i in $m^3\,kg^{-1}\,s^{-2} \times 10^{-13}$.

	i							
	1	2	3	4	5	6	7	8
x_i	6.674	6.6729	6.6709	6.674 255	6.675 59	6.674 22	6.674 07	6.673 87
s_i	0.07	0.05	0.07	0.009	0.003	0.098	0.033	0.027

Table 3. Gas concentration data for seven labs in $\mu mol\,mol^{-1}$ units.

	i						
	1	2	3	4	5	6	7
x_i	9.961	9.979	10.012	10.013	10.026	10.038	10.495
s_i	0.205	0.174	0.078	0.086	0.158	0.063	0.503

be too short, and may not achieve the nominal coverage. In practice the considered estimators provide, on average, similar values for the consensus value. It is their uncertainties which heavily depend on the estimation method.

Acknowledgment

The author is grateful to Antonio Possolo for his advice and help.

References

[1] Mandel J 1991 *Evaluation and Control of Measurements* (New York: Dekker)
[2] Crowder M 1992 Interlaboratory comparisons: round robins with random effects *Appl. Stat.* **41** 409–25
[3] Cox D R 1982 Combination of data *Encyclopedia of Statistical Sciences* vol 2 ed S Kotz and N L Johnson (New York: Wiley) pp 46–53
[4] Nielsen L 2000 Evaluation of measurement intercomparisons by the method of least squares *Technical Report* DFM-99-R39 Danish Institute of Fundamental Metrology
[5] Cox M G 2002 The evaluation of key comparsion data *Metrologia* **39** 589–95
[6] Zhang N F 2006 The uncertainty associated with the weighted mean of measurement data *Metrologia* **43** 195–204
[7] Pinheiro J and Bates D 2000 *Mixed Effects Models in S and S-Plus* (New York: Springer)
[8] Pinheiro J, Bates D, DebRoy S and Sarkar D 2007 *nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-83* the R Core team
[9] Vangel M G and Rukhin A L 1999 Maximum likelihood analysis for heteroscedastic one-way random effects ANOVA in interlaboratory studies *Biometrics* **55** 129–36
[10] Graybill F A and Deal R B 1959 Combining unbiased estimators *Biometrics* **15** 543–50
[11] Mohr P J and Taylor B N 2000 CODATA recommended values of the fundamental physical constants: 1998 *Rev. Mod. Phys.* **72** 351–495
[12] Voinov V G and Nikulin M S 1993 *Unbiased Estimators and Their Applications* (Dordrecht: Kluwer)
[13] Fairweather W R 1972 A method for obtaining an exact confidence interval for the common mean of several normal populations *Appl. Stat.* **21** 229–33
[14] Willink R 2002 Statistical determination of a comparison reference value using hidden errors *Metrologia* **39** 343–54
[15] Cochran W G 1937 Problems arising in the analysis of a series of similar experiments *J. R. Stat. Soc.* Supplement **4** 102–18
[16] Rao P S R S 1981 Cochran's contributions to variance component models for combining estimates *W G Cochran's Impact on Statistics* ed P Rao and J Sedransk (New York: Wiley)
[17] Rao P S R S, Kaplan J and Cochran W G 1981 Estimators for the one-way random effects model with unequal error variances *J. Am. Stat. Assoc.* **76** 89–97
[18] Chunovkina A, Elster C, Lira I and Wöger W 2008 Analysis of key comparsion data and laboratory biases *Metrologia* **45** 211–6
[19] Duewer D L 2008 A comparison of location estimators for interlaboratory data contaminated with value and uncertainty outliers *Accred. Qual. Assur.* **13** 193–216
[20] Chunovkina A and Cox M G 2003 A model-based approach to key comparsion data evaluations *Proc. 17th IMEKO World Congress (Dubrovnik, Croatia)*
[21] Rukhin A L 2003 Two procedures of meta-analysis in clinical trials and interlaboratory studies *Tatra M. Math. Publ.* **28** 155–68
[22] Kacker R 2004 Combining information from interlaboratory evaluations using a random effects model *Metrologia* **41** 132–6
[23] DerSimonian R and Laird N 1986 Meta-analysis in clinical trials *Control. Clin. Trials* **7** 177–88
[24] Mandel J and Paule R C 1970 Interlaboratory evaluation of a material with unequal number of replicates *Anal. Chem.* **42** 1194–7
[25] Paule R C and Mandel J 1982 Consensus values and weighting factors *J. Res. Natl Bureau Stand.* **87** 377–85
[26] Rukhin A L, Biggerstaff B and Vangel M G 2000 Restricted maximum likelihood estimation of a common mean and Mandel–Paule algorithm *J. Stat. Plan. Inference* **83** 319–30
[27] Schiller S and Eberhardt K 1992 Combining data from independent chemical analysis methods *Spectrochim. Acta* **12** 1607–13
[28] Rukhin A L and Vangel M G 1998 Estimation of a common mean and weighted means statistics *J. Am. Stat. Assoc.* **93** 303–8
[29] Rukhin A L 2007 Estimating common vector mean in interlaboratory studies *J. Multivariate Anal.* **98** 435–54

[30] Cox D R 1975 A note on partially Bayes inference and the linear model *Biometrika* **62** 651–4
[31] Horn R A, Horn S A and Duncan D B 1975 Estimating heteroscedastic variance in linear models *J. Am. Stat. Assoc.* **70** 380–5
[32] Rukhin A L 2007 Conservative confidence intervals based on weighted means statistics *Stat. Probab. Lett.* **77** 1312–21
[33] Dietrich C F 1991 *Uncertainty, Calibration and Probability: The Statistics of Scientific and Industrial Measurement* 2nd edn (Bristol: A Hilger)
[34] ISO 1995 *Guide to the Expression of Uncertainty in Measurement* 2nd edn (Geneva: International Organization for Standardization)
[35] Rukhin A L and Sedransk N 2007 Statistics in metrology: international key comparisons and interlaboratory studies *J. Data Sci.* **7** 393–412
[36] Mohr P J and Taylor B N 2005 CODATA recommended values of the fundamental physical constants: 2002 *Rev. Mod. Phys.* **77** 1–108
[37] Guenther F *et al* 2007 International Comparison CCQM-K41: Hydrogen sulfide in nitrogen *Metrologia* **44** (Tech. Suppl.) 08004