Interlaboratory consensus building

Interlaboratory consensus building

G Mana

INRIM – Istituto Nazionale di Ricerca Metrologica, Torino, Italy
UNITO – Università di Torino, Dipartimento di Fisica, Torino, Italy

E-mail: g.mana@inrim.it

Abstract. Determining the degree of equivalence of participating laboratories from the results of measurement comparisons still prompts discussions among metrologists, especially where measurement uncertainties have been underestimated. This paper expands on a solution to a problem issued in 2020 by the Journal of Analytical and Bioanalytical Chemistry. The example illustrates an approach to consensus-building based on Bayesian selection among statistical models that attempt to explain the excess of data variation. The probability of any further model being correct can be similarly calculated.

Submitted to: *Metrologia*

PACS numbers: 02.50.Cw, 02.50.-r, 02.50.Tt, 07.05.Kf

1. Introduction

When measurement results are inconsistent, that is, their dispersion is greater than what is expected from the reported uncertainties and allowing for different statistical models, determining a reference value of the measurand prompts discussions [1–8]. A typical case study is to determine a consensus value of the Newtonian constant of gravitation [9]. Reviews of the state-of-the-art in the selection of statistical models and data reduction in laboratory comparisons are in [10, 11] and the references therein.

An excess of data scatter suggests the presence of unrecognized contributions to the uncertainty budget, colloquially referred to as *dark uncertainty* [10,12]. The problem is to cope with the this scatter, which is larger than the reported uncertainties. An approach is to inflate the individual uncertainties [13]. Another is to identify and exclude from the calculation of the reference value the outliers. This exclusion occurs naturally if the data are sampled from distributions with thick tails, e.g., a Student t-distribution [14–18]. An alternative to distributions with thick tails is a "good-and-bad data" model, which leads to a mixture of two Gaussians [17,19]. Another, by assuming normal data whose means are shifted by unknown-variance zero-mean laboratory effects, is to use a random-effect model [10, 20].

A decision-theoretical approach requires to assign probabilities to the measurand values. Since they follow from the application of the probability calculus, which is an extension of the Boolean logic, these probabilities encode in a consistent way the information available on the measurand value before the measurements are carried out and that delivered by the measurement results. They underpin any informed decision based on the measurand value, for instance, the choice of the reference value. Eventually, their normalisation factor makes it possible to compare the statistical models proposed for the scatter of the data, to identify which model is the best explanation of the discrepancy, and to choose the reference value based on it.

The solution presented here extends a Bayesian approach outlined in [17,18]. To illustrate it, I consider the challenge in [21, 22], which addresses determining the reference value and the degree of equivalence of the laboratories participating in a comparison of activity measurements of the radionuclide ^{59}Fe [23].

Since the chi-squared test for the variance detects mutual inconsistency and sample preparation or artefact instability is ruled out, I assume that the uncertainties associated with the data might be only lower bounds to standard deviations and develop the tools necessary to i) compare statistical models that groups differently the data and ii) select the most probable model, given the data. This makes it unnecessary to identify and exclude the data disagreeing with the majority or magnify the uncertainties to make the data consistent.

Calculations were carried out with the aids of Mathematica [24]. The relevant notebook is given as supplementary material. To read and interact with it, download the Wolfram Player free of charge from Wolfram Research. Reference books on the probability calculi and Bayesian inferences underpinning this paper are [25–30].

2. Problem statement

The interlaboratory comparison considered involves eleven metrology institutes [23]. Following the challenge in [21], the input data (see [21] and the supplementary material) are the measured values of the iron-59 activity, x_i, which is positive by definition, and the associated uncertainties, u_i, which are judged underestimated by a chi-squared test for the variance. No a priori knowledge is assumed about correlations, degrees of freedom of the uncertainty estimate, and the measurand.

The first step in Bayesian inferences is to encode the information available on the measurement procedure in a probability density function of the input data [31]. The distributions that encode only that the data are unbiased measurement results of the same quantity having specified uncertainties, without introducing uncontrolled assumptions, are independent Gaussians [28, 32], having common (positive) mean μ and, to avoid neglecting underestimation, standard deviations σ_i greater than or equal to the uncertainties associated to the measurement results.

We might also explain the data scatter by noting that the uncertainties are measurement results [17, 33] and by allowing for standard deviations either wider or narrower. I do not examine this possibility, but its greater or lesser probability can be determined along the same lines discussed in the following.

3. Solution

Data inconsistency implies that there are unrecognised contributions to the uncertainty budget [12] or, which is equivalent, that the uncertainties associated with the data are lower bounds to standard deviations [17, 18]. No additional information is available about the data dispersion, e.g., about artefact instability or the sample preparation. Therefore, to explain the data, I consider the following statistical models. For some datum, the $\sigma_i = u_i$ identity holds; the others are affected by unrecognized uncertainty contributions.

In the first case, $x_i \sim N(x_i|\mu, u_i^2)$. In the second, $x_i \sim N(x_i|\mu, \sigma_i^2)$, where $\sigma_i \geq u_i$. The hypothesis space contains 2048 (mutually exclusive) models, classified by the subsets of the measured values, the empty set and its complement included. Each subset groups the results $x_i \sim N(x_i|\mu, u_i^2)$, whose associated uncertainties are the standard deviations of their sampling distributions.

Since any model is uncertain, to prove or disprove that the chosen one explains the data, I must allow comparisons against the others. This requirement imposes that the marginal likelihood (the normalisation factor in the Bayes' rule, also termed evidence) is independent of the model parameterisations. This because it is proportional to the probability of observing the data no matter what the model parameters may be. Consequently, the prior distributions of different model parameterisations must be proper and comply with the change-of-variable rule.

Since testable information is not given, the area element of the $N(x_i|\mu, \sigma_i^2)$ manifold equipped with a Fisher-information metric (in general, named after H Jeffreys [34]) does the work without introducing uncontrolled assumptions. Hence, the prior distribution of the model parameters μ and σ_i is

$$\mu, \sigma_i \sim \pi(\mu, \sigma_i|u_i) = \frac{u_i}{V_\mu \sigma_i^2}, \tag{1}$$

where $u_i \leq \sigma_i$, $\mu > 0$, and V_μ is the μ's support.

The main motivation for this choice is that it complies with the change-of-variable rule under reparameterisation. Further motivations, built on the symmetries of the way the measurand is linked to the data, are discussed in [30]. The supplementary material shows explicitly that the parameterisations $N(x_i|\mu, u_i^2 + \tau_i^2)$ or $N(x_i|\mu, \lambda^2 u_i^2)$ deliver the same prior after the $\sigma_i^2 = u_i^2 + \tau_i^2$ and $\sigma = \lambda u$ changes of variable. A different prior is possible [20], but we must be aware of the information delivered to the problem and accept that the changing-of-variable rule will give the distribution of any other parameterisation.

When an improper prior is used for the parameters, the posterior probability of the model being correct is meaningless [35]. Therefore, without affecting the proportionality to the area element and the compliance with the change-of-variable rule, (1) restricts the range of the μ values to a finite interval, which is assumed large enough to allow approximating the needed integrations by extending V_μ to the real line. At the same time, keeping V_μ explicit makes (1) dimensionally correct and future comparisons possible.

The sampling distribution of x_i, given μ and u_i and with the unknown σ_i integrated out, is

$$\begin{aligned} x \sim L(x|\mu, u) &= u \int_u^{+\infty} N(x|\mu, \sigma^2)/\sigma^2 \, \mathrm{d}\sigma \\ &= \frac{\left(1 - \mathrm{e}^{-\frac{(x-\mu)^2}{2u^2}}\right) u}{\sqrt{2\pi}(x-\mu)^2}, \end{aligned} \tag{2}$$

where I dropped the i subscript. The data likelihood is

$$\mathbf{x} \sim Q(\mathbf{x}|\mu, \mathbf{u}, A) = \prod_{i \in A} N(x_i|\mu, u_i^2) \prod_{j \in \bar{A}} L(x_j|\mu, u_j), \tag{3}$$

where A is a subset grouping $x_i \sim N(x_i|\mu, u_i^2)$ data and $\bar{A}$ is its complement. The marginal likelihood and the posterior distribution of the mean are

$$Z(\mathbf{x}|\mathbf{u}, A) = \frac{1}{V_\mu} \int_{-\infty}^{+\infty} Q(\mathbf{x}|\mu, \mathbf{u}, A) \, \mathrm{d}\mu \tag{4}$$

and

$$\mu \sim P(\mu|\mathbf{x}, \mathbf{u}, A) = \frac{Q(\mathbf{x}|\mu, \mathbf{u}, A)}{V_\mu Z(\mathbf{x}|\mathbf{u}, A)}, \tag{5}$$

where the V_μ/u_i support and $\mu/u_i > 0$ are large enough to extend the integration to the real line for all practical purposes. The integral in (4) has no analytical solution. As shown in the supplementary material, I evaluated it numerically with the help of Mathematica.

The probabilities of the models A_i, $i = 1, 2, \ldots 2048$, of being correct when the measurement results are $\{x_i\}$, which are shown in Fig. 1, are

$$\mathrm{Prob}(\mathrm{A_i}|\mathbf{x}, \mathbf{u}) = \frac{\mathrm{Z}(\mathbf{x}|\mathbf{u}, \mathrm{A_i})}{\sum_\mathrm{i} \mathrm{Z}(\mathbf{x}|\mathbf{u}, \mathrm{A_i})}, \tag{6}$$

where, in the absence of additional information, I assumed equiprobable A_is, which corresponds to the maximum entropy prior.

4. Consensus value

All the information about the measurand is encoded in its posterior probability density (5) averaged over all the models,

$$\mu \sim \sum_i P(\mu|\mathbf{x}, \mathbf{u}, A_i) \mathrm{Prob}(\mathrm{A_i}|\mathbf{x}, \mathbf{u}). \tag{7}$$

For the sake of simplicity, I pick up the most probable model, A_mx (see Figs. 1 and 2). Hence,

$$\mu \sim P(\mu|\mathbf{x}, \mathbf{u}, A_\mathrm{mx}) \tag{8}$$

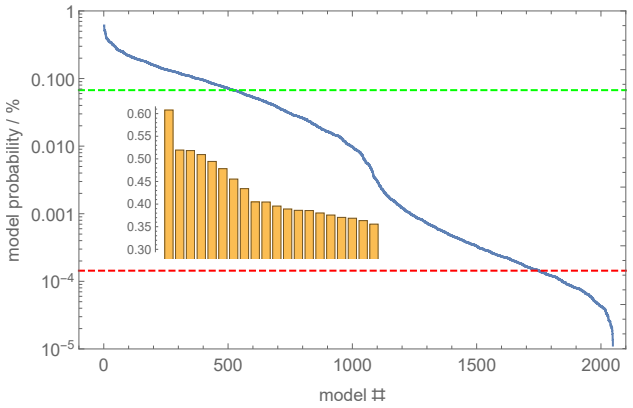

Figure 1. Posterior probabilities, sorted in decreasing order, of the 2048 data subsets to group the measurement results whose associated uncertainties are equal to the standard deviation (blue line), see Eq. (6). The inset shows the first 20 values. The horizontal lines are the posterior probabilities of all $\sigma_i > u_i$ (green) and all $\sigma_i = u_i$ (red) subsets.

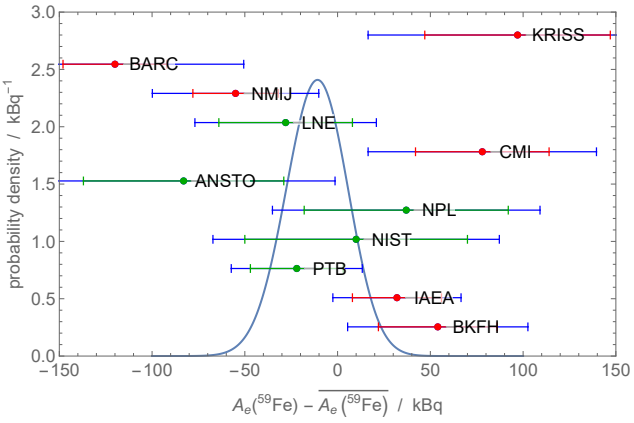

Figure 2. Most probable posterior probability, see Eq. 5, density for the activity of iron-59. $\overline{A_e(^{59}\mathrm{Fe})} = 14,631$ kBq is the arithmetic mean of the data. The dots are the measured values; the bars are the associated uncertainties u_i (green: $\sigma_i = u_i$, red: $\sigma_i > u_i$) and posterior 68% credible intervals (blue).

and

$$Z(\mathbf{x}|\mathbf{u}, A_{\mathrm{mx}}) = (61 \times 10^{-27}\,\mathrm{kBq}^{-10})/\mathrm{V}_\mu. \qquad (9)$$

To explain the data, other models are possible. Therefore, the A_{mx}'s evidence (9) lets competing explanations be compared with A_{mx}. It is a kindness to those who may wish to check this explanation against alternatives without having to redo the calculations. As an example, the evidence of the no-Gaussian-datum model is $(6.7 \times 10^{-27}\,\mathrm{kBq}^{-10})/\mathrm{V}_\mu$, whereas that of the all-Gaussian-data one is $(0.01 \times 10^{-27}\,\mathrm{kBq}^{-10})/\mathrm{V}_\mu$ (see the supplementary material).

The choice of a consensus value is a matter of minimisation of the agreed cost function. The posterior mean, mode, and median are all equal to $\overline{\mu} = 14,620$ kBq. The posterior standard deviation is 16 kBq. The interval that, with 95% probability, includes the true activity is $[14,588, 14,652]$ kBq (see the supplementary material). For a comparison, the Possolo's consensus value is 14,628 kBq, with a 95% interval $[14,585, 14,674]$ kBq [22].

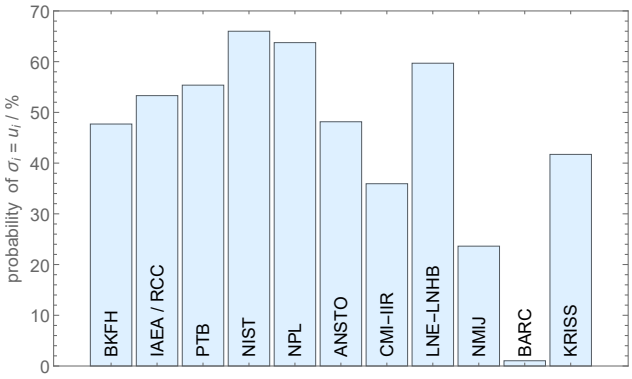

Figure 3. Posterior probability that the standard deviation of the measurement result is equal to the associated uncertainty, see Eq. (10).

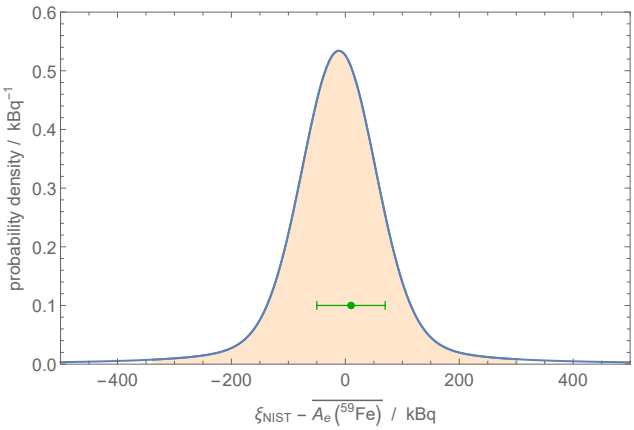

Figure 4. Predictive distribution of future NIST measurement results, see Eq. (12), given the data explanation A_{mx}. The filled area is the 95% confidence-interval. The dot is NIST's result; the bar is the associated uncertainty.

5. Degrees of equivalence

The goal of a laboratory comparison is to assess the *validity of the accuracy of national measurement standards and of calibration and measurement certificates* [1]. The gauges are the deviations from the reference value determined from the comparison and from one another, as expressed by the unilateral and bilateral degrees of equivalence. The unilateral degree of equivalence is the deviation of each result from the consensus value and the associated expanded uncertainty for the 95% confidence level. The degree of equivalence between pairs of results is the difference of their deviations from the consensus value and the associated expanded uncertainty for the 95% confidence level. I

do not go on in these lines but investigate alternative ways to assess the participants' capability.

5.1. First option.

The first way is to evaluate the probability that no unrecognised term contributed to the uncertainty, which is the same probability that the given uncertainty equals the standard deviation. Since A_is are mutually exclusive, this probability (see Fig. 3) is

$$\mathrm{Prob}(\sigma_k = u_k) = p_k = \sum_{i:x_k \in A_i} \mathrm{Prob}(\mathrm{A_i}|\mathbf{x}, \mathbf{u}), \qquad (10)$$

For instance, the probability that the standard deviation of the NIST's measurement result is equal to the associated uncertainty is 66% (see the supplementary material).

5.2. Second option.

The second way uses the predictive distribution of the future measurement result ξ of the k-th laboratory, which is the mixture

$$H(\xi|\mu, \mathbf{x}, \mathbf{u}) = p_k N(\xi|\mu, u_k^2) + (1 - p_k) L(\xi|\mu, u_k^2), \quad (11)$$

where the p_k probability of the $N(\xi|\mu, u_k^2)$ model is given by (10). The relevant 68% credible intervals are shown in Fig. 2.

Given the A_{mx} model and marginalising over the posterior distribution of μ, the predictive distribution of the measurement result is

$$\xi|\mathbf{x}, \mathbf{u}, A_{\mathrm{mx}} \sim \int_{-\infty}^{+\infty} H(\xi|\mu, \mathbf{x}, \mathbf{u}) P(\mu|\mathbf{x}, \mathbf{u}, A_{\mathrm{mx}})\, \mathrm{d}\mu. \quad (12)$$

Figure 4 shows the probability density (12) of an additional NIST's measurement result, whatever the measurand value might be. The predicted mean is still $\overline{\mu} = 14,620$ kBq. The interval that, with 95% probability, includes the future NIST results is $[14,293, 14,946]$ kBq (for the computation, see the supplementary material). This interval is nearly three times wider than that expected from the sampling distribution, $N(\xi|\mu, u_{\mathrm{NIST}}^2)$, associated with the NIST result; that is, 653 kBq *vs.* 235 kBq. This difference is due to both the μ uncertainty and the residual 34% probability of missing terms in the uncertainty budget.

6. Conclusions

The probabilities assigned to the measurand values consistent with the data and information at hands offer a way to agree on a measurand value in laboratory comparisons, no matter whether the results are consistent or not and without worry about outliers.

When there are competing statistical data models, the model most supported by data can be identified by calculating each one's probability of being true. Next, the posterior distribution of the measurand values can be agreed according to the most likely or the marginalised one.

I illustrated this approach by applying it to the results of activity measurements of the radionuclide ^{59}Fe, which are inconsistent due to underestimated uncertainties. Any further model competing to explain the data can be easily integrated into the analysis and assessed based on the reported evidence.

Acknowledgments

Support was received from the Ministero dell'Istruzione, dell'Università e della Ricerca. Thanks to the referees for their careful reading and the many valuable suggestions that helped improve the manuscript clarity.

References

[1] Comité International des Poids et Mesures 1999 Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates, Technical Supplement revised in October 2003 Sèvres: BIPM URL `https://www.bipm.org/en/cipm-mra/cipm-mra-documents`
[2] Cox M G 2002 *Metrologia* **39** 589–595
[3] Cox M G 2007 *Metrologia* **44** 187–200
[4] Lira I 2007 *Metrologia* **44** 415–421
[5] Toman B 2007 *Technometrics* **49** 81–87
[6] Toman B 2009 *Accred. Qual. Assur.* **14** 553–563
[7] Elster C and Toman B 2010 *Metrologia* **47** 113–119
[8] Forbes A B 2016 *Metrologia* **53** 1295–1305
[9] Merkatas C, Toman B, Possolo A and Schlamminger S 2019 *Metrologia* **56** 054001
[10] Koepke A, Lafarge T, Possolo A and Toman B 2017 *Metrologia* **54** S34–S62
[11] Possolo A, Koepke A, Newton D and Winchester M 2021 *J Res Natl Inst Stan* **126** 126007
[12] Thompson M and Ellison S 2011 *Accred. Qual. Assur.* **16** 483–487
[13] Birge R T 1932 *Phys. Rev.* **40**(2) 207–227
[14] O'Hagan A 1979 *Journal of the Royal Statistical Society. Series B (Methodological)* **41** 358–367
[15] Hanson K M and Wolf D R 1996 Estimators for the Cauchy distribution *Maximum Entropy and Bayesian Methods* ed Heidbreder G R (Dordrecht: Springer) pp 255–263
[16] Hanson K M 2005 *AIP Conference Proceedings* **803** 431–439
[17] Sivia D and Skilling J 2006 *Data Analysis: A Bayesian Tutorial* (Oxford: Oxford University Press)
[18] Mana G, Massa E and Predescu M 2012 *Metrologia* **49** 492–500
[19] Press W H 1997 Understanding data better with bayesian and global statistical methods *Unsolved Problems in Astrophysics* ed Bahcall J N and Ostriker J P (Princeton: Princeton University Press) pp 49–60
[20] Bodnar O, Muhumuza R N and Possolo A 2020 *Metrologia* **57** 064004
[21] Possolo A 2020 *Anal. Bioanal. Chem.* **412** 3955–3956
[22] Possolo A 2021 *Anal. Bioanal. Chem.* **413** 3–5
[23] Michotte C, Ratel G, Courte S, Kossert K, Nähle O, Dersch R, Branger T, Bobin C, Yunoki A and Sato Y 2019 *Metrologia* **57** 06003
[24] Wolfram Research, Inc 2020 Mathematica, Version 12.2 champaign, IL URL `https://www.wolfram.com/mathematica`

[25] Jaynes E and Bretthorst G 2003 *Probability Theory: The Logic of Science* (Cambridge: Cambridge University Press)
[26] D'Agostini G 2003 *Bayesian Reasoning in Data Analysis: A Critical Introduction* (Singapore: World Scientific)
[27] MacKay D 2003 *Information Theory, Inference and Learning Algorithms* (Cambridge: Cambridge University Press)
[28] Gregory P 2005 *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica® Support* (Cambridge: Cambridge University Press)
[29] von der Linden W, Dose V and von Toussaint U 2014 *Bayesian Probability Theory: Applications in the Physical Sciences* (Cambridge: Cambridge University Press)
[30] Harney H 2016 *Bayesian Inference: Data Evaluation and Decisions* (Switzerland: Springer International Publishing)
[31] Joint Committee for Guides in Metrology 2008 Evaluation of measurement data – Supplement 1 to the "Guide to the expression of uncertainty in measurement" – Propagation of distributions using a Monte Carlo method Sèvres: BIPM URL `https://www.bipm.org/en/publications/guides`
[32] Mana G and Pizzocaro M 2021 *Metrologia* **58** 015012
[33] Hanson K M and Wolf D R 1999 Outlier tolerant parameter estimation *Maximum Entropy and Bayesian Methods* ed von der Linden W, Dose V, Fischer R and Preuss R (Dordrecht: Springer) pp 47–56
[34] Robert C P, Chopin N and Rousseau J 2009 *Statistical Science* **24** 141 – 172
[35] Bodnar O and Eriksson V 2021 Bayesian model selection: Application to adjustment of fundamental physical constants (*Preprint* `2104.01977`)