Statistical interpretation of key comparison degrees of equivalence based on distributions of belief

2007 Metrologia 44 L14

(http://iopscience.iop.org/0026-1394/44/2/N03)

IOP PUBLISHING METROLOGIA

Metrologia **44** (2007) L14–L17 doi:10.1088/0026-1394/44/2/N03

SHORT COMMUNICATION

Statistical interpretation of key comparison degrees of equivalence based on distributions of belief

Blaza Toman

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

Received 16 January 2007
Published 27 March 2007
Online at stacks.iop.org/Met/44/L14

Abstract
The Mutual Recognition Arrangement (MRA) requires the calculation of unilateral degrees of equivalence (DoE), defined as the difference between a laboratory's measurement and a key comparison reference value (KCRV). The functional form of the KCRV has been a source of much debate, with most of the proposed solutions being based on a frequentist interpretation of probability. The focus of this article is on the case when the result of each laboratory's measurement is interpreted as a mean and standard deviation of a probability distribution of belief about a measurand. Bayesian statistical methods are best suited to deal with this situation and lead to interesting insights about the form of the KCRV and the DoE.

1. Introduction

Recent publications of the Joint Committee for Guides in Metrology (JCGM), such as for example the Supplement 1 to the *Guide to the Expression of Uncertainty in Measurement* [1], make it clear that the interpretation of probability used in uncertainty quantification in metrology is to be consistent with Bayesian statistics. This then makes it necessary to use methods of key comparison analysis which have clear Bayesian interpretation. There is now a growing literature addressing this problem; some recent examples are [2–5]. These papers focus almost entirely on the problem of computing the key comparison reference value (KCRV) and its uncertainty, not the degrees of equivalence (DoE). The definition given in the Technical Supplement to the arrangement [6] is simple: the DoE of each laboratory is defined to be the deviation from the KCRV. What is not totally straightforward, however, is the computation of the uncertainty of this deviation, also required by [6]. This is not clearly defined and can be interpreted in various ways. This paper considers the simple case when each participating laboratory measures the value X of a quantity associated with some stable artefact [7] and addresses specifically the problem of calculation of the uncertainty of the DoE. Adopting the interpretation of probability consistent with Bayesian statistics, the results from laboratory i, that is, x_i and the uncertainty $u_i(x)$, are interpreted as the mean and standard deviation of a probability distribution $f_i(x)$. This describes the laboratory's belief about X. If there are p participants in the key comparison, there are p belief distributions about X. In the same context, articles [2,3] propose methods of computing a single reference value for the measurand. In the process, several methods of combining probability distributions that have appeared in the statistics literature, for example in [8], are considered. In [3], the method of combining the p distributions is the logarithmic pool method. A different approach to combining probability distributions is employed by [2], one called the Supra Bayesian method. This method is further extended in this paper to specifically address the problem of calculation of the DoEs and their uncertainties. Section 2 presents the general model and formulae. Section 3 uses a hypothetical example to illustrate the results and to compare them with results based on the usual weighted mean analysis. Conclusions follow in section 4

2. The Supra Bayesian analysis

The Supra Bayesian (SB) method, traditionally a method of combining expert judgments, has appeared in many publications, for example in [9, 10]. In the current context, it can be described as follows. A single decision maker (the pilot laboratory) consults p experts (laboratories) who provide the means x_i and standard deviations $u_i(x)$ of their probability distributions for X. The decision maker then combines the p experts' distributions into a single probability distribution for X. He does this by specifying a likelihood function for the data (the x_i), by specifying a prior distribution for X and any other parameters of the model, and then applying Bayes' theorem.

The problem addressed here is that of the calculation of the unilateral DoE and the accompanying uncertainty. As the DoE are in effect estimates of laboratory biases, it is necessary to include these in the likelihood function in order to obtain a probability distribution for them. (A similar point was made in the frequentist context by [11].) Matrix notation simplifies the definition of the likelihood function. Let $\underline{x} = \left(x_1, \ldots, x_p\right)'$ and

$$\Sigma_1 = \begin{bmatrix} u_1^2(x) & 0 & 0 & 0 \\ 0 & . & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & u_p^2(x) \end{bmatrix}.$$

Further let

$$A = \begin{bmatrix} 1 & 1 & 0 & . & . & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & . & 0 & 0 \\ 1 & 0 & 0 & 0 & . & 0 \\ 1 & 0 & . & . & 0 & 1 \end{bmatrix}$$

that is, a $p \times (p+1)$ matrix with the first column being all ones and the remaining columns forming a $p \times p$ identity matrix. Also define $\underline{\theta}_1 = (X, \alpha_1, \ldots, \alpha_p)'$, where the α_i are the laboratory biases. Then the pilot's likelihood function for $\underline{x}$ can be taken as a Gaussian distribution with mean $A \cdot \underline{\theta}_1$ and covariance matrix Σ_1. To apply Bayes' theorem, it is necessary to define a prior distribution for $\underline{\theta}_1$. Let $\underline{\theta}_2 = (X_0, 0, \ldots, 0)'$, a vector of size $p+1$, where X_0 is any constant. Further define a $(p+1) \times (p+1)$ diagonal matrix

$$\Sigma_2 = \begin{bmatrix} \tau_0 & 0 & 0 & 0 & 0 \\ 0 & \tau_1 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & \tau_1 \end{bmatrix}$$

where $\tau_i > 0$ for $i = 0, 1$. Finally, define the prior distribution of $\underline{\theta}_1$ to be

$$\underline{\theta}_1 \sim N(\underline{\theta}_2, \Sigma_2).$$

Application of Bayes' theorem, as shown in [12], leads to a posterior distribution for $\underline{\theta}_1$, that is

$$\underline{\theta}_1 | \underline{x} \sim N\left(\left[A' \cdot \Sigma_1^{-1} \cdot A + \Sigma_2^{-1}\right]^{-1} \left(A' \cdot \Sigma_1^{-1} \underline{x} + \Sigma_2^{-1} \underline{\theta}_2\right), \right.$$
$$\left. \left[A' \cdot \Sigma_1^{-1} \cdot A + \Sigma_2^{-1}\right]^{-1}\right). \qquad (1)$$

As previously pointed out, for example in [3], the pilot laboratory does not generally possess any information about the α_i. Further, in the key comparison context, it is necessary that the analysis be as objective as possible. This is best achieved by the use of prior distributions which are essentially non-informative. In the above model this is accomplished by letting $\tau_0 \to \infty$ and $\tau_1 \to \infty$. The resulting marginal distribution of X is Gaussian with mean

$$\mu_X = \overline{x} = \frac{1}{p} \sum_{i=1}^{p} x_i \qquad (2)$$

and standard deviation

$$\omega_X = \frac{1}{p} \sqrt{\sum_{i=1}^{p} u_i^2 (x)}. \qquad (3)$$

The posterior distribution of the α_i is also Gaussian with mean

$$\mu_{\alpha_i} = x_i - \overline{x} \qquad (4)$$

and standard deviation

$$\omega_{\alpha_i} = \sqrt{u_i^2 (x) + \frac{1}{p^2} \sum_{j=1}^{p} u_j^2 (x) - \frac{2}{p} u_i^2 (x)}. \qquad (5)$$

Thus the resulting KCRV is $\overline{x}$, and the DoE for laboratory i, that is, $x_i - \overline{x}$ is in fact the posterior mean of the laboratory bias, a natural estimate of α_i. Thus expression (5) is clearly the accompanying uncertainty. This set of estimates is called the SB solution in the next section.

The above model, under the assumptions that $\tau_0 \to \infty$ and $\tau_1 \to 0$ (the laboratory biases are essentially equal to 0), obtains the usual weighted mean solution [7], that is, the posterior distribution of X is Gaussian with mean

$$\mu_{\mathrm{WM}} = \frac{\sum_1^p x_i u_i^{-2} (x)}{\sum_1^p u_i^{-2} (x)} \qquad (6)$$

and standard deviation

$$\omega_{\mathrm{WM}} = \frac{1}{\sqrt{\sum_1^p u_i^{-2} (x)}}. \qquad (7)$$

In the next section, a hypothetical example is used to compare the performance of the SB solution with that of the weighted mean.

3. Example

Consider a fictitious key comparison experiment with p laboratories where $(p-1)$ of the laboratories have identical results (same x_i) and one laboratory differs from the others. Table 1 gives the details of the data.

This is an extreme case of a situation which is not uncommon; that is, of one possibly outlying laboratory with a comparatively small uncertainty. This example points out some key differences between the SB procedure and that given in [7]. The SB solutions are

$$\mu_X = \frac{a\,(p-1+k)}{p}, \quad \omega_X = \frac{1}{p}\sqrt{(p-1)+b},$$
$$\mu_{\alpha_i} = \frac{(1-k)\,a}{p},$$
$$\omega_{\alpha_i} = \sqrt{\left(1 - \frac{2}{p}\right) + \frac{1}{p^2}\,(p-1+b)}, \quad i = 1, \ldots, p-1,$$

Table 1. Data for a hypothetical key comparison experiment in arbitrary units.

Laboratory	Estimate x_i	Uncertainty $u_i(x)$
$1 \ldots (p-1)$	a	1
p	ka	b

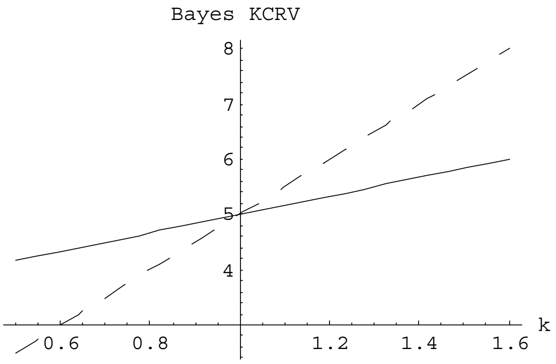

Figure 1. The SB solution KCRV μ_X (solid line) as a function of k. The dashed line is the value of x_i for the outlying laboratory 3.

$$\mu_{\alpha_i} = \frac{(p-1)(k-1)a}{p},$$
$$\omega_{\alpha_i} = \sqrt{\left(1-\frac{2}{p}\right)b + \frac{1}{p^2}(p-1+b)}, \quad i = p. \tag{8}$$

These can be compared with the results for the weighted mean analysis, that is

$$\mu_{\text{WM}} = \frac{a(p-1+k/b)}{(p-1+1/b)}, \qquad \omega_{\text{WM}} = \sqrt{\frac{1}{(p-1)+1/b}},$$
$$\text{doe}_i = \frac{(1/b)(1-k)a}{(p-1+1/b)},$$
$$u(\text{doe}_i) = 1 - \frac{1}{(p-1)+1/b} \quad i = 1, \ldots, p-1,$$
$$\text{doe}_i = \frac{(p-1)(k-1)a}{(p-1+1/b)},$$
$$u(\text{doe}_i) = b - \frac{1}{(p-1)+1/b} \quad i = p. \tag{9}$$

To observe the behaviour of the estimators in (8) and (9), consider the case when $a = 5$, $b = 0.25$ and $p = 3$. This is a case of comparison of three laboratories with one laboratory having a standard uncertainty of one fourth of the other two. It is an example that is relatively extreme and thus clearly illustrates the differences between the two solutions. Figures 1 and 2 show the behaviour of the two KCRV estimators as a function of k. It is clear that μ_X is less sensitive to change in k, in other words, more robust to outliers.

Usually, in key comparison experiments, the main interest lies in the 95% probability intervals of the DoEs. If a particular interval includes 0 then the laboratory appears to agree with the others in the key comparison. Given the expressions in (8) and (9) it is possible to observe when this happens as a function of the values of k. Figure 3 shows that the range of values of k for which laboratories 1 and 2 appear to be in good

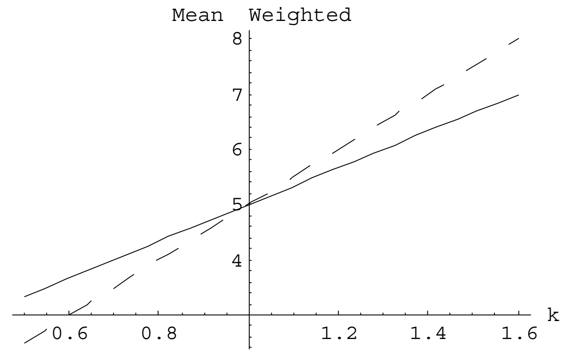

Figure 2. The weighted mean KCRV μ_{WM}(solid line) as a function of k. The dashed line is the value of x_i for the outlying laboratory 3.

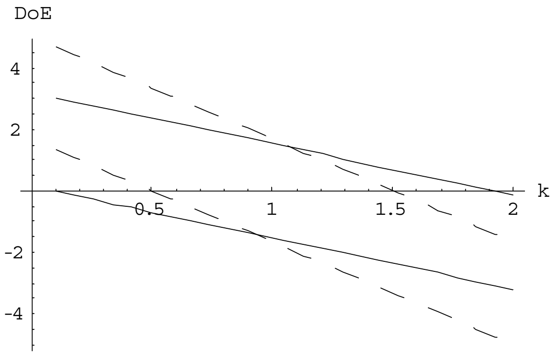

Figure 3. The solid lines are the lower and upper bounds of a 95% coverage interval for the SB solution α_i, for $i = 1, 2$. The dashed lines are the lower and upper bounds of a 95% coverage interval for the weighted mean based DoE.

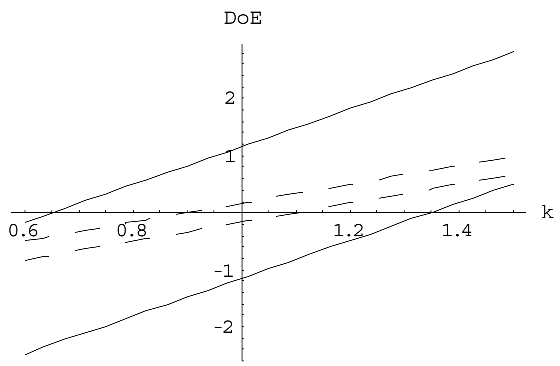

Figure 4. Lower and upper bounds of a 95% coverage interval for α_i, for $i = 3$. The dashed lines are the lower and upper bounds of a 95% coverage interval for the weighted mean based DoE.

agreement with the entire group (this includes laboratory 3) is approximately (0.1, 1.9) for the Bayesian method, and (0.5, 1.5) for the weighted mean method. Thus the SB procedure is also more robust to the effects of an outlying laboratory when determining DoE. For laboratory 3, the same behaviour, even more extreme, is illustrated in figure 4. Laboratory 3 will appear in good agreement when k is in the interval (0.65, 1.35) for the SB method, but only when k is in the interval (0.9, 1.1) for the weighted mean method. Thus it is again clear that the DoE based on the weighted mean are much less robust to outlying laboratory results than the estimates based on the SB model presented here.

4. Conclusions

This paper presents a method of estimation of unilateral degrees of equivalence and the accompanying uncertainty in a key comparison experiment. This method uses the belief-based interpretation of probability, one consistent with the *Guide to the Expression of Uncertainty in Measurement* [13]. In addition, it is shown here by example that the method produces results that are generally more robust to outlying observations than the usual weighted mean method.

References

[1] http://www.intermet.jp/JCGM/0610news/Supplement to GUM.pdf
[2] Toman B 2007 Bayesian approaches to calculating a reference value in key comparison experiments *Technometrics* **49** 81–7
[3] Willink R 2006 Forming a comparison reference value from different distributions of belief *Metrologia* **43** 12–20
[4] Dose V 2007 Bayesian estimate of the Newtonian constant of gravitation *Meas. Sci. Technol.* **18** 176–82
[5] Lira I 2006 Bayesian evaluation of comparison data *Metrologia* **43** S231–4
[6] http://www.bipm.org/utils/en/pdf/mra_techsuppl2003.pdf
[7] Cox M G 2002 The evaluation of key comparison data *Metrologia* **39** 589–95
[8] Genest C and Zidek J V 1986 Combining probability distributions *Stat. Sci.* **1** 114–48
[9] Keeney R L and Raiffa H 1976 *Decisions with Multiple Objectives: Preferences and Value Tradeoffs* (New York: Wiley)
[10] Lindley D 1983 Reconciliation of probability distributions *Oper. Res.* **31** 866–80
[11] Kacker R N, Datla R U and Parr A C 2003 Statistical interpretation of key comparison reference value and degrees of equivalence *J. Res. Natl Inst. Stand. Technol.* **108** 439–46
[12] Lindley D and Smith A F M 1972 Bayes estimates for the linear model *J. R. Stat. Soc.* B **34** 1–41
[13] ISO Technical Advisory Group (Working Group 3) 1993 *Guide to the Expression of Uncertainty in Measurement* (Geneva: International Organization for Standardization)