Computational Statistics and Data Analysis 55 (2011) 1815–1827

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Laplace random effects models for interlaboratory studies

Andrew L. Rukhin *, Antonio Possolo

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

ARTICLE INFO

Article history:
Received 15 June 2010
Received in revised form 15 November 2010
Accepted 16 November 2010
Available online 25 November 2010

Keywords:
Bayes estimators
Degree of equivalence
Heteroscedasticity
Interlaboratory study
Key comparisons
Laplace distribution
Maximum likelihood estimation
Meta-analysis
Random effects model

ABSTRACT

A model is introduced for measurements obtained in collaborative interlaboratory studies, comprising measurement errors and random laboratory effects that have Laplace distributions, possibly with heterogeneous, laboratory-specific variances. Estimators are suggested for the common median and for its standard deviation. We provide predictors of the laboratory effects, and of their pairwise differences, along with the standard errors of these predictors. Explicit formulas are given for all estimators, whose sampling performance is assessed in a Monte Carlo simulation study.

Published by Elsevier B.V.

1. Interlaboratory studies and key comparisons

The international agreement, the so-called "Mutual Recognition Arrangement" (MRA) (CIPM, 1999) on mutual recognition of national measurement standards, calibration and measurement certificates issued by national metrology institutes (NMIs) calls for the execution of interlaboratory studies aimed at testing principal techniques and measurement methods in a particular field of science. These studies are organized by the Consultative Committees (CCs) of the *Comité International des Poids et Mesures* (CIPM), there are CCs for length, mass, amount of substance, etc. Such interlaboratory studies are called *Key Comparisons* (KCs), and one of their principal goals is to establish the degree of equivalence of national measurement standards which characterize the extent to which each institute may have confidence in the results reported by other NMIs. Typically, a KC produces a *key comparison reference value* (KCRV): for example, in a KC focusing on the length of a gauge block, this should be the block's true length (ISO/IEC, 2007, 5.18) although in actuality it is the best estimate of this length. In a KC focusing on the mass fraction of a particular substance in a certified reference material of which aliquots are distributed to the participating NMIs for analysis, this could be the mass fraction of one or more selected compounds.

The MRA defines the *Degree of Equivalence* of a national measurement standard (unilateral DoE) as comprising its deviation from the key comparison reference value and the uncertainty of this deviation. According to the MRA the degree of equivalence between a pair of national measurement standards (bilateral DoE) is formed by the difference of their deviations from the reference value and the uncertainty of this difference. If a reference value cannot be meaningfully defined (for example, when the KC involves multiple circulating artifacts and not all NMIs measure all of them), the KC results might be expressed directly in terms of the degrees of equivalence between pairs of standards.

* Corresponding author. Fax: +1 301 975 3144.
E-mail addresses: andrew.rukhin@nist.gov (A.L. Rukhin), antonio.possolo@nist.gov (A. Possolo).

0167-9473/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.csda.2010.11.016

The *International vocabulary of metrology* (VIM) (ISO/IEC, 2007, 2.26) defines *measurement uncertainty* as a "non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used", and adds that this "parameter may be, for example, a standard deviation". For this reason we follow the customary usage in statistics and use either "standard deviation" or "standard error" throughout, where metrologists might use "measurement uncertainty" instead.

The ith of n NMI participating in a KC is supposed to produce a measured value x_i and an assessment of its standard error u_i. Thus we start with a set of n (scalar) measurement values $x_1, \ldots, x_n$, and the corresponding standard errors $u_1, \ldots, u_n$, which we assume are known. The precise meaning of these uncertainties often is debated heatedly. Could they be regarded as known quantities, or instead are they merely estimates of unknown quantities? Some metrologists insist that uncertainties are *computed* (with assuredness and certainty of arithmetic), while others concede that they are only *estimated*. When x_i and u_i are modeled as in Bayesian inference, one can convincingly argue that conditionally upon the data, u_i indeed are computed as standard deviations of suitable posterior distributions, and therefore are known with certainty which is our assumption.

The next section introduces the model. We find estimators of the KCRV (8) and of its standard error (10), as well as of a scale parameter (9) in the distribution of the random interlaboratory effects. Assuming these parameters to be known, in Section 3 the estimators of the degrees of equivalence (value and standard error) are derived. Section 4 contains results of Monte Carlo simulation. Most of the formulas needed in Section 3 are collected in the Appendix.

2. Random effects and common median model

2.1. Mixed effects Laplace model

In many practical cases, all confidence intervals based on measured values x_i and their standard errors u_i do not overlap, which suggests that the dispersion of the measured values x_i is greater than what their standard errors might lead one to expect. It is also fairly common that a few of the measurements deviate markedly from the bulk of the rest.

The first kind of situation can be dealt with by modeling the measurements as outcomes of independent random variables $X_i = \mu + B_i + E_i$ for $i = 1, \ldots, n$, where μ is the unknown KCRV, B_i is a lab-specific random effect, and E_i represents measurement error. Similarly to a common practice in robust estimation (Wilcox, 2005) the second eventuality can be addressed by modeling the distributions of B_i and E_i as suitably heavy-tailed. The results are then analyzed using either *ad hoc* robust statistical methods, or likelihood methods, conventional or Bayesian, that guarantee similar robustness within a parametric framework.

There is a precedent to this general approach. For example, Pinheiro et al. (2001) describe a model that is based on Student's t-distribution. In the same spirit, the median has been suggested as a possible consensus KCRV estimator e.g., Cox (2002). However, this method does not use the standard errors u_i at all; neither do other robust estimators that have been suggested to address the same problem, e.g., Analytical Methods Committee (1989a,b) and Thompson et al. (2006). The departures from the Gaussian random effects linear model, that appear most detrimental to the performance of the estimators, while staying within the realm of symmetric distributions, are heaviness of the tails of the distribution of B_i, and incomplete knowledge of the variances of the measurement errors E_i.

To account for these facts we suggest a mixed effects model in which both B_i and E_i have suitably scaled Laplace (double-exponential) distributions. This model, as we shall show, is far more robust than the traditional Gaussian model, while incurring only moderate loss of efficiency in this traditional case. Similarly to what other laboratory effects do, it also overcomes the problem of "inconsistency" (between the x_i). See Toman and Possolo (2009) for critique of consistency testing proposed by Decker et al. (2006) and Cox (2007). In our model both lab-specific random effects and within-lab measurement errors can be interpreted as Gaussian but with variances u_i^2 that are like random draws from exponential distributions.

Our estimation method for μ on the basis of heterogeneous data is based on the statistic,

$$\widetilde{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \frac{|x_i - \mu|}{u_i}, \tag{1}$$

which is a weighted median. This procedure has a maximum likelihood interpretation to be discussed in the next section, and there are efficient numerical algorithms for its evaluation; see Bloomfield and Steiger (1983). Indeed, medians weighed by their standard errors have already been suggested as KCRV estimators; see Müller (2000), Ratel (2006). Besides those mentioned above, Rocke (1983), Davies (1991), Lischer (1996) and Duewer (2006, 2008) advocate the use of robust statistics (including the median) in interlaboratory studies. The same robustness issues arise in the more general context of meta-analysis; see Hedges and Olkin (1985). To address meta-analysis problems robustly, Demidenko (2004) uses a setting which is somewhat similar to the following model by assuming Gaussian errors and Laplace between-lab effects. Wilcox (2006) points out difficulties with homogeneity testing for the medians.

The model we propose for KCRV estimation, and for the assessment of its standard error is this: the measured values, $x_1, \ldots, x_n$, are outcomes of random variables

$$X_i = \mu + B_i + E_i, \tag{2}$$