ARTICLE IN PRESS
Measurement xxx (2016) xxx–xxx
Contents lists available at ScienceDirect
Measurement
journal homepage: www.elsevier.com/locate/measurement

Proficiency tests with uncertainty information: Detection of an unknown random effect

Katsuhiro Shirono *, Masanori Shiro, Hideyuki Tanaka, Kensei Ehara

National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 3058565, Japan

ARTICLE INFO

Article history:
Received 7 May 2015
Received in revised form 9 October 2015
Accepted 4 January 2016
Available online xxxx

Keywords:
Model selection
Empirical Bayes
Marginal likelihood
Interlaboratory comparison
ISO 13528

ABSTRACT

In proficiency tests (PTs), the possible existence of an unknown random effect (such as instability or inhomogeneity of the measured items, or vagueness in the definition of the measurand) that could affect the reported values is a matter of concern, and may influence the performance evaluations in an unfair manner. If an unknown random effect is the dominant source of uncertainty, it is not appropriate to conduct performance evaluations without correcting for that random effect (increasing the uncertainties, correcting the biases or the both). This study presents a statistical method to detect an unknown random effect before the performance evaluation in a PT with uncertainty information. The method is validated through simulations using various types of virtual but possible data sets. Through the application of this method, the applicability of the PT data to the performance evaluation can be checked.

1. Introduction

A proficiency test (PT) by means of an interlaboratory comparison is an effective tool to assure the quality of the measurements of calibration and testing laboratories [1]. In ISO/IEC 17025 [2], participation in a PT is stipulated as one of the methods of validating calibration and measurement capability. For performance evaluation in a PT with uncertainty information, a comparison with the result of a reference laboratory is usually implemented. Statistical methods for use in such a case are given in ISO 13528 [3].

It is sometimes difficult, however, to identify an appropriate reference laboratory for certain PTs, including key comparison tests [4] conducted among national metrology institutes. A guideline for the analysis of key comparison test results has been provided by Cox [5]. In this guideline, the consistency of the data is checked by means of the χ^2 test. For inconsistent data, the guideline provides a performance evaluation method in which a statistical model is not explicitly given. Other than this guideline, analysis using the largest consistent subset (LCS) proposed by Cox [6] has often been employed for key comparison tests. The LCS is the subset with the largest data size among the subsets whose consistencies are confirmed through the χ^2 test. In the statistical model in this method, no reliabilities are given for measurement results that are outside of the LCS. This type of analysis is referred to as LCS analysis in this paper.

It should be noted that the cause of inconsistency is not investigated in these analysis methods. Two main causes of inconsistency can be considered to exist: (i) an unknown random effect that affects most of the reported data (such as instability or inhomogeneity of the measured items, or vagueness in the definition of the measurand), and (ii) the presence of one or more unskilled laboratories that report outlier values or underestimated uncertainties. It is, in fact, possible that both of these may be the causes

* Corresponding author.
E-mail addresses: k.shirono@aist.go.jp (K. Shirono), shiro@ni.aist.go.jp (M. Shiro), tanaka-hideyuki@aist.go.jp (H. Tanaka), ehara.kensei@aist.go.jp (K. Ehara).

http://dx.doi.org/10.1016/j.measurement.2016.01.002
0263-2241/

Please cite this article in press as: K. Shirono et al., Proficiency tests with uncertainty information: Detection of an unknown random effect, Measurement (2016), http://dx.doi.org/10.1016/j.measurement.2016.01.002

of an inconsistency. The statistical model in LCS analysis is appropriate when the second cause is dominant.

When an unknown random effect is one of the causes of an inconsistency, the performance evaluation should not be conducted without correction of the reported data. For example, when the random effect is essentially caused by instability of the measured item, the reported values may be randomly determined by such instability irrespective of the proficiencies of the laboratories. Moreover, the same can be said when the random effect is essentially due to vagueness in the definition. Correction is therefore necessary before the performance evaluation. Distinctively appropriate corrections (increasing uncertainties or correcting biases) have been proposed in previously reported studies using a random effect model [7–9]. In this paper, analysis using a statistical model with a random effect is referred to as random effect model (REM) analysis.

Despite the importance of elucidating the causes of inconsistency, a statistical tool to detect an unknown random effect in an actual PT has not yet been made available as long as we know. Although there has been some statistical studies to detect and quantify an unknown random effect [10,11], only highly skilled participants are considered in most of them. In other words, a methodology to select the best statistical model from among an REM and models in which the random effect is not considered when outliers exist has not been offered so far. Our research group, for example, has reported a number of statistical models for the analysis of key comparison tests [12], but has not given a criterion for model selection. Therefore, the reasons for the occurrence of an inconsistency need to be discussed from a technical standpoint.

This study provides a method to detect a random effect in the data of a PT in which uncertainty information is given, through selection of the model using the marginal likelihood [13]. The marginal likelihood is, to put it simply, the likelihood of the statistical model. If the marginal likelihood of the REM is larger than that of the model without the random effect, it can be concluded that an unknown random effect has been detected. If a random effect has been detected, the reported data are inapplicable to the performance evaluation.

Although various methods using Bayesian statistics have been reported for the analysis of PTs [14–26], only a few of these apply the model selection approach. Our research group has reported some results of the analysis of PT data using the marginal likelihood [14,15], but the REM was not taken into consideration in these studies. In particular, it was shown in Ref. [15] that the analysis results might sometimes be unreliable if the REM is not taken into consideration. Mana et al. [16] have also reported the analysis of some studies for determination of the Planck constant using the model selection approach including the REM. It could be said that the model selection in the study by Mana et al. is extended in the present study to provide greater flexibility. Such flexibility is important for robust analysis in the presence of outliers. Moreover, the Bayesian model averaging technique is employed in some previous studies [17,18].

Performance evaluation is not the main interest of this study and discussed in another paper [27]. The computational procedure presented there is given in Section 3.2 of the present paper. One reason why this performance evaluation method is necessary is that the results given by the proposed method are occasionally different from those obtained by LCS analysis. The method proposed in this study places importance on the measured values with a small uncertainty relative to the LCS analysis, as described in Section 4. We believe that the proposed analysis method will enhance the motivation of the participants to report not overestimated but pertinent uncertainty due to this difference.

This paper is organized as follows: The statistical models are given in Section 2. Section 3 presents practical procedures for both cases in which a random effect is either detected or undetected. In Section 4, the validity of the proposed method is confirmed through the application to numerical PT data and the properties are compared with those of LCS analysis and REM analysis. The contents of the paper are then briefly summed up in Section 5. The Appendices A and B provide an explanation of the marginal likelihood employed in this study and the computation, respectively. More information on the computation is given in the electronic supplementary materials (ESMs). The ESMs consists of a PDF file and a TXT file of the Microsoft® Visual Basic® for Applications (VBA7) source code for Microsoft® Excel® 2013.

2. Statistical model

2.1. Proposed statistical model

A statistical model to detect the random effect is presented in this section. Let n be the number of participants. The following parameters are the hyperparameters of the likelihood and the priors, and are determined to maximize the marginal likelihood:

1. Number of data to which the common random effect is applicable: m. ($m = 0, 2, 3, \ldots, n$).
2. Laboratory identification number $K(i)$, where $i = 1, 2, \ldots, n$. ($K(1) < K(2) < \cdots < K(m)$, $K(m+1) < K(m+2) < \cdots < K(n)$).
3. Parameters for the prior: $\alpha, \beta_{m+1}, \beta_{m+2}, \ldots$, and β_n. ($1 \leqslant \alpha < +\infty$, $1 \leqslant \beta_i$ ($i = 1, 2, \ldots, n$)).

Suppose that Laboratory $K(i)$ reports the measurement value x_i and its standard uncertainty u_i ($i = 1, 2, \ldots, n$). Let $q_i = u_i^2$ for simplicity of the description. x_i is assumed to be derived from the normal distribution with the same mean of μ. On the other hand, the variances of the distribution for the reported values of Laboratories $K(1), K(2), \ldots, K(m)$ are assumed to be $q_i + \theta_c$, where θ_c is the variance caused by the unknown random effect. The variances for the reported values of Laboratories $K(m+1), K(m+2), \ldots, K(n)$ are assumed to be $q_i + \theta_i$, where θ_i is the variance caused by the unskillfulness of the laboratory. The statistical model for x_i is then given as follows:

$$\begin{aligned} x_i &\sim \mathrm{N}(\mu, q_i + \theta_c) \quad \text{for } i = K(1), K(2), \ldots, K(m), \\ x_i &\sim \mathrm{N}(\mu, q_i + \theta_i) \quad \text{for } i = K(m+1), K(m+2), \ldots, K(n). \end{aligned} \tag{1}$$