Unilateral degree of equivalence maximizing the power of test in an analysis of a regional metrology organization key comparison

Katsuhiro Shirono¹, Maurice Cox²
1 National Metrology Institute of Japan, Japan
2 National Physical Laboratory, UK
k.shirono@aist.go.jp

Abstract. Data analysis of a regional metrology organization (RMO) when linked to an International Committee of Weights and Measures key comparison (KCs) is discussed. One of the purposes of an RMO KC is to check the appropriateness of participants’ results from a statistical point of view. For this purpose, a statistical test is useful. In this study, we discuss the derivation of a unilateral Degree of Equivalence (DoE) as a test statistic, and apply the test to dummy data. The unilateral DoEs given by the method in this study can be significantly different from those given by previous studies. These differences can be ascribed to the use of different statistical models. A Consultative Committee (CC) should choose an appropriate statistical model for the analysis in accordance with its purpose in implementing an RMO KC.

1. Introduction

The mutual recognition arrangement (MRA) [1] has played an essential role in realizing global metrological traceability. The Calibration and Measurement Capability (CMC) of a National Metrology Institute (NMI) is established through the framework of the MRA. Key comparisons (KCs) support the CMCs from a technical point of view. While KCs are often conducted by the International Committee of Weights and Measures (CIPM), Regional Metrology Organizations (RMOs) implement KCs for the same purpose, because a CIPM KC is not open to all NMIs. It is described in the MRA [1] that "participation in a CIPM key comparison is open to laboratories having the highest technical competence and experience, normally the member laboratories of the appropriate Consultative Committee". Moreover, it is explicitly required in Technical Supplement T.4 of the MRA that "the results of the RMO key comparisons are linked to key comparison reference values established by CIPM key comparisons by the common participation of some institutes in both CIPM and RMO comparisons". Thus, the linking of an RMO KC to the corresponding CIPM KC is a realistic way to establish metrological traceability for the RMO laboratories that participate. Laboratories that participate in both CIPM and RMO KCs are called linking laboratories in this study.

In both types of KC, the performance of a laboratory is evaluated using a unilateral Degree of Equivalence (DoE). The unilateral DoE consists of a value part and an uncertainty part. The value part is the difference between the reported and the reference values, and the uncertainty part is the uncertainty associated with this difference at a 95 % level of confidence [1]. The unilateral DoE quantifies the performance of a laboratory. Through the linking of two KCs, it is expected to quantify the relationship between the unilateral DoEs in these KCs. However, no official guidance about the linking procedure has been provided by the CIPM, despite the fact that there are several suggestions from the academic point of view [2]-[6].

One reason why there are a number of linking methods is that several statistical models can be considered to establish the unilateral DoE including the statistical model we employ in this study, which is given as follows. Suppose m and n laboratories participate in a CIPM KC and a corresponding RMO KC, respectively. Further suppose there are L linking laboratories, for which $x_i$ and $y_i$ are the reported values from laboratory $i$ in the CIPM and RMO KCs, respectively ($i = 1, \ldots, L$). $x_{L+1}$ to $x_m$ are the reported values from non-linking laboratories $L+1$ to $m$ in the CIPM KC. Similarly, $y_{L+1}$ to $y_n$ are those from non-linking laboratories $L+1$ to $n$ in the RMO KC. For $i > L$, $x_i$ and $y_i$ are
reported by different laboratories. The following statistical model, assuming underlying normality, is proposed:

\[
\begin{align*}
    x_i &\sim N(\mu_i, u^2(x_i)) \quad \text{for } i = 1, \ldots, m, \\
y_i &\sim N(\mu_i, u^2(y_i)) \quad \text{for } i = 1, \ldots, n, \\
    \rho_j &\equiv \frac{\text{cov}(x_j, y_j)}{u(x_j)u(y_j)} \quad \text{for } i = 1, \ldots, L,
\end{align*}
\]

where \(\mu_i\) and \(\mu_j\) are unknown values of the measurands in the CIPM and the RMO KCs, respectively. Further, \(u(x_i)\) and \(u(y_j)\) are the standard uncertainties associated with \(x_i\) and \(y_j\), respectively, and \(\rho_j\) and \(\text{cov}(x_j, y_j)\) are respectively the correlation coefficient and the covariance associated with \(x_j\) and \(y_j\) for the linking laboratories. Values for \(\text{cov}(x_j, y_j)\) and hence \(\rho_j\) can typically be obtained from the uncertainty budgets for the linking laboratories by identifying the effects that are common to the measurements made in both the CIPM and the RMO comparisons as implemented in some previous studies [7]. We call model (1) the non-bias model, because biases are not considered.

In general, even if the statistical model is given, there are still some candidates for the unilateral DoE in an analysis of RMO KCs. In other words, there are several mathematical expressions for the unilateral DoE with which we can test the non-bias model. In this study, we focus on the power of the test. Namely, the optimal mathematical expression of the unilateral DoE is suggested so that the power of the test is maximized.

2. Maximization of the power of the test

A particular non-linking laboratory of concern is given the index \(j (j > L)\). Since we determine the power of the test for laboratory \(j\) with respect to a parameter \(\theta_j\), model (1) is extended to incorporate the parameter as follows:

\[
\begin{align*}
    x_i &\sim N(\mu_i, u^2(x_i)) \quad \text{for } i = 1, \ldots, m, \\
y_i &\sim N(\mu_i, u^2(y_i)) \quad \text{for } i = L + 1, \ldots, j - 1, j + 1, \ldots, n, \\
y_j &\sim N(\mu_j + \theta_j, u^2(y_j)) \quad \text{for } i = L + 1, \ldots, j - 1, j + 1, \ldots, n, \\
    \rho_j &\equiv \frac{\text{cov}(x_j, y_j)}{u(x_j)u(y_j)} \quad \text{for } i = 1, \ldots, L.
\end{align*}
\]

Here, \(\theta_j\) is the parameter defining the power of the test, \(p(\theta_j)\), with respect to laboratory \(j\). (We show subsequently that the mathematical expression of the statistic to maximize the power is independent of \(\theta_j\) and hence of the non-linking laboratories.) Defining the null and alternative hypotheses as \(H_0\) and \(H_1\), respectively, the power of the test is the probability that \(H_0\) is rejected when \(H_1\) is true. In this case, \(H_0\) and \(H_1\) are given as follows:

\[
\begin{align*}
    H_0 : \theta_j &= 0, \\
    H_1 : \theta_j &\neq 0.
\end{align*}
\]

Only linear combinations of \(x_i\) and \(y_i\) with the following mathematical formula are considered in this study to test \(H_0\):

\[
    e_j = \sum_{i=1}^{m} g_{i,j} x_i + \sum_{i=1}^{n} h_{i,j} y_i.
\]

Since the probability distributions underlying the \(x_i\) and \(y_i\) have been assumed normal, the probability distribution of \(e_j\) is also normal, with mean \(E[e_j]\) and variance \(V[e_j]\) given as follows:

\[
    E[e_j] = \sum_{i=1}^{m} g_{i,j} \mu_i + \sum_{i=1}^{n} h_{i,j} \mu_i + h_{i,j} \theta_j,
\]

\[
    V[e_j] = g' \Sigma g,
\]

where \(g_j = (g_{1,j}, h_{1,j}, \ldots, g_{m,j}, h_{m,j}, g_{1,j}+1,h, \ldots, g_{m,j}+1,h, \ldots, h_{m,j}+1,h)\), and \(\Sigma\) is the covariance matrix of the vector \(x_j = (x_1, y_1, \ldots, x_i, y_i, x_{i+1}, \ldots, x_n, y_n, \ldots, y_n)\). Vector \(g_j\) is optimized to maximize the power of the test. When the significance level is set as 5%, the power of the test is

\[
p(\theta_j) = 1 - \Phi\left(1.96 - \frac{e_j - E[e_j]}{V[e_j]}\right) + \Phi\left(-1.96 - \frac{e_j - E[e_j]}{V[e_j]}\right)
\]

where \(\Phi(\cdot)\) is the cumulative probability function of the standardized normal distribution.
Since the vector $g_i$ is not uniquely determined, we impose the following (unrestrictive) conditions:

\[ E[\varepsilon_i] = h_j \theta_j, \]
\[ V[\varepsilon_i] = 1, \]
\[ h_j > 0. \]  

(7)

Fortunately, we can determine the vector $g_i$ that maximizes the power of the test in formula (6) subject to condition (7) irrespective of the value of $\theta_j$. With the optimum vector $g_i$, $\varepsilon_i$ is given as follows:

\[ \varepsilon_i = \frac{y_j - y_{ref}}{u^2(y_j) - u^2(y_{ref})}, \]  

(8)

where $y_{ref}$ denotes the best linear estimate of $\mu_j$ based on model (1). Note that the numerator and denominator in the right-hand side of (8) are consistent with formulae (3) and (5) in Cox [8]. $y_{ref}$ is specifically given as follows:

\[ y_{ref} = \frac{1}{u^2(y_{\text{link}})} \sum_{i=1}^{n} \frac{1}{u^2(y_i)} \left[ \frac{y_{\text{link}}}{u^2(y_{\text{link}})} + \sum_{i=1}^{n} \frac{y_i}{u^2(y_i)} \right]. \]  

(9)

Here, $y_{\text{link}}$ and $u^2(y_{\text{link}})$ are given as follows:

\[ y_{\text{link}} = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} \Sigma_i^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} \Sigma_i^{-1} z_i \end{pmatrix}, \]  

(10)

\[ u^2(y_{\text{link}}) = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} \Sigma_i^{-1} \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \]

We employ the subscript “ref” for $y_{ref}$ because it can be considered as a candidate for the reference value of the RMO KC. It should be noted that the variance of $y_j - y_{ref}$ is given as $u^2(y_j) - u^2(y_{ref})$ so that $V[\varepsilon_i] = 1$.

Thus, we suggest the value and uncertainty parts of the unilateral DoE are as follows:

\[ d_j = y_j - y_{ref}, \]
\[ U_j = 1.96 \times \sqrt{u^2(y_j) - u^2(y_{ref})}. \]  

(12)

because $d_j$ in formula (12) is intuitively understandable when $y_{ref}$ is actually the reference value, and, as the $E_u$ score in ISO/IEC 17043 [9], we can define $E_u = d_j/U_j$, which is an appropriate index to be employed in the statistical test in terms of its power.

3. Example

The proposed unilateral DoE is applied to a dummy example. Table 1 shows the dummy data for a CIPM KC and a corresponding RMO KC. The results of applying the method of this paper are shown in Table 2. In a typical case, the performance is evaluated to be satisfactory when $|d_j/U_j|$ is no greater than one. It is found that the performance of all non-linking laboratories (laboratories 2 to 4) is satisfactory with the method we propose. If another evaluation method is applied, different results are obtained. For example, the results from the method proposed by Elster et al. [6] are also shown in Table 2. Since the values of $|d_j/U_j|$ exceed 1 for laboratories 3 and 4, the performance of these laboratories is assessed to be unsatisfactory. It should be noted that the statistical model (implicitly) considered in that study is a model with biases, and different from the model we consider. The differences shown in this example imply that an appropriate statistical model should be chosen in accordance with the purpose of a comparison. When the quantification of existing biases is of interest, estimation using a model with biases is required. However, when we wish to assess whether a participant’s result has bias or not, the proposed statistical test with the non-bias model is a reasonable way to proceed.
Table 1 – Dummy data of a CIPM KC and an RMO KC. Laboratory 1 is the sole linking laboratory. No correlation between $x_i$ and $y_i$ is assumed.

| Laboratory | $x_i$ | $u(x_i)$ | Laboratory | $y_i$ | $u(y_i)$ |
|------------|------|-------|------------|------|-------|
| 1 (linking lab.) | 0.0 | 0.5 | 1 (linking lab.) | 0.0 | 0.5 |
| 2          | -1.3 | 1.0 | 2          | 1.7 | 1.0 |
| 3          | -1.3 | 1.0 | 3          | 1.8 | 1.0 |
| 4          | -1.3 | 1.0 | 4          | 1.9 | 1.0 |
| 5          | -1.3 | 1.0 | 5          | 1.0 | 1.0 |

Table 2 – Computed unilateral DoE for the dummy data in Table 1 by the method of this study and that of Elster et al. [6]

| Laboratory | Proposed unilateral DoE | Unilateral DoE according to [6] |
|------------|------------------------|--------------------------------|
|            | $d_i$ | $U_i$ | $d_i/U_i$ | $d_i$ | $U_i$ | $d_i/U_i$ |
| 2          | 0.9  | 1.8  | 0.5       | 2.4  | 2.3  | 1.0       |
| 3          | 1.0  | 1.8  | 0.6       | 2.5  | 2.3  | 1.1       |
| 4          | 1.1  | 1.8  | 0.6       | 2.6  | 2.3  | 1.1       |

4. Summary
For the performance evaluation of a laboratory in an RMO KC, a statistical test using the unilateral DoE is considered. The unilateral DoE maximizing the power of test is proposed based on a non-bias model. The models employed in some previous studies are not all the same as the model we suggest. While an appropriate model may vary depending on the purpose of a KC, it can be said that our proposal is the best tool in terms of the power of the test when the purpose of an RMO KC is to detect the existence of a bias rather than to quantify bias.

Acknowledgement
The work of the first author was partly supported by JSPS KAKENHI Grant Number 17K18411. The National Measurement Office of the UK’s Department for Department for Business, Energy and Industrial Strategy supported the second author and in part the first author as part of NPL’s Data Science programme.

References
[1] Bureau International des Poids et Mesures. Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, 1999.
[2] D. R. White. On the analysis of measurement comparisons. Metrologia, 41(3):122, 2004.
[3] C. M. Sutton. Analysis and linking of international measurement comparisons. Metrologia, 41(4):272, 2004.
[4] I. A. Kharitonov and A. G. Chunovkina. Evaluation of regional key comparison data: two approaches for data processing. Metrologia, 43(5):470, 2006.
[5] J. E. Decker, A. G. Steele, and R. J. Douglas. Measurement science and the linking of CIPM and regional key comparisons. Metrologia, 45(2):223, 2008.
[6] C. Elster, A. G. Chunovkina, and W. Wöger. Linking of a RMO key comparison to a related CIPM key comparison using the degrees of equivalence of the linking laboratories. Metrologia, 47(1):96, 2010.
[7] M. G. Cox, P. M. Harris. The evaluation of key comparison data using key comparison reference curves. Metrologia, 49(4):437, 2012.
[8] M. G. Cox. The evaluation of key comparison data. Metrologia, 39:589–595, 2002.
[9] International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), ISO/IEC 17043: conformity assessment – general requirements for proficiency testing, ISO, Geneva, 2010.