XVII IMEKO World Congress
Metrology in the 3rd Millennium
June 22 - 27, 2003, Dubrovnik, Croatia

A MODEL-BASED APPROACH TO KEY COMPARISON DATA EVALUATION

*Anna Chunovkina**, *Maurice Cox***

* D.I.Mendeleyev Institute for Metrology, St. Petersburg, Russia
** National Physical Laboratory, Teddington, UK

Abstract – The evaluation of key comparison data is discussed, the general case of correlated data being considered. Particular attention is paid to a simplified procedure for data evaluation, founded on a mixture of distributions associated with the results from the institutes participating in the comparison. The suggested approach uses the model of an interlaboratory experiment from ISO 5725 and uncertainty evaluation in accordance with the Guide to the Expression of Uncertainty in Measurement.

Keywords: key comparison, reference value, degree of equivalence, mixture distribution.

1. INTRODUCTION

The mutual recognition of national measurement standards and of calibration and measurement certificates is an essential feature of international cooperation for quality assurance of measurements. A main objective of the MRA [1] is “to provide … a secure technical foundation for wider agreements related to international trade, commerce and regulatory affairs”. The MRA is realized through key comparisons (KCs) and supplementary comparisons, and through the quality systems of the national metrological institutes (NMIs). The MRA poses the following problems:

- Establishing the degrees of equivalence (DoEs) of measurement standards on the basis of KC data
- Confirmation of the assigned NMI capabilities on the basis of KC data
- Linking the results of Regional Metrology Organization (RMO) comparisons to Comité International des Poids et Mesures (CIPM) KCs.

In this paper statistical methods for the solution of these problems are considered.

The choice of the data evaluation algorithm should be dictated by information relating to the measurement data. The statistical methods used are based on assumptions for which relevant valid information should be made available. Without such information the quality of the results obtained cannot be determined. Statisticians working in metrology have an obligation to provide a framework for the correct application of these methods. Data evaluation guidelines should explain the conditions of use, which should be formulated so as to avoid misleading interpretations.

The BIPM (Bureau International des Poids et Mesures) Director's Advisory Group on Uncertainty has developed some guidelines for KC data evaluation [2], in which two of the above problems are addressed when the measurements are mutually independent. However, correlation between data often occurs in practice, for reasons of traceability to a common source, the use of the same measurement method, etc. So, handling correlated data is important, as is the linking of RMO and CIPM KCs, which have their own correlation structure through the NMIs in both comparisons.

If the data covariance matrix is known *a priori* (or quantifiable from valid information), a straightforward extension of the guidelines [2] permits correlation to be treated. However, often such information is not available. Quantifying the covariance matrix requires detailed analysis of the sources of uncertainty for each participating NMI. Even after such an analysis, which requires the joint efforts of experts from these NMIs, doubts concerning the quantitative expressions of correlation can remain. So other approaches to handle correlated data are of interest. The main attention in the paper is given to a simplified method for handling correlated data when inadequate information is readily available to quantify the covariance matrix.

Section 2 presents the rationale of the approaches discussed. Section 3 considers a model based on the principles of ISO 5725 and the use of a mixture distribution to summarize the KC data. Section 4 is concerned with data analysis and using these models. Section 5 discusses the linking of KCs. Section 6 indicates some alternative approaches. Section 7 contains concluding remarks.

2. RATIONALE

The DoE [1] is the degree to which the NMIs’ standards are consistent with reference values determined from the KCs and hence are consistent with one another. The DoE of each measurement standard is expressed quantitatively by two terms: its deviation from the KC reference value (KCRV) and the uncertainty of this deviation (at a 95% level of confidence):

$$\{d_i, 2u(d_i)\}, \qquad d_i = x_i - x_{\mathrm{ref}}, \tag{1}$$

where x_i is the value provided by the *i*th NMI, and x_{ref} denotes the KCRV. The DoE of national standards is interpreted as the equivalence of participating NMIs' measurements. For a pair of standards the DoE is expressed as the difference between the results obtained and the uncertainty of this difference:

$$\{d_{i,j}, 2u(d_{i,j})\}, \qquad d_{i,i} = x_i - x_j. \tag{2}$$

The main data evaluation problem is establishing the KCRV, for which the valid application of mathematical and statistical methods is required to avoid subjectivism. Often the weighted mean, with weights equal to the inverse squares of the standard uncertainties associated with the NMIs' results is used as the KCRV, an estimate having the smallest uncertainty under the following conditions:

1. The standard uncertainties can be regarded as valid
2. The value provided are mutually independent
3. A Gaussian distribution can be assigned to the measurand realized by each participating NMI.

For periodical comparisons the Working Group of the respective Consultative Committee of the CIPM can agree the relevant uncertainty components and their manner of evaluation. Doing so helps to confirm the validity of the combined standard uncertainties for the participating NMIs.

For a new KC, when the aims are to establish the reproducibility of results and reveal the reasons for the differences between them, condition 1 is less likely to apply. The only information that can arguably be taken as valid is the characteristics of reproducibility for each NMI estimated from a large number of measurements. In terms of the provided uncertainties, if not all the corresponding coverage intervals "overlap sufficiently", the implication is that some of these uncertainties are underestimated and hence invalid.

In quantifying the covariance matrix, according to the GUM a probabilistic approach should be used to treat both random and systematic effects. Random effects are independent within and between the NMIs. Reasons for the mutual dependence of systematic effects in different NMIs include common traceability, similar equipment, and similar realization of the measurement procedure. Because it can be difficult to quantify the covariance matrix, covariances are often taken as zero. However, doing so might result in an unjustified simplification of the data evaluation problem, including the use of an inadequate model for the KC data.

The assignment of a Gaussian distribution to an NMI's measurand may be attributed to an application of the Central Limit Theorem or to the fact that only the measurement result and the associated uncertainty are provided, no other information on the form of distribution being available.

Since the weighted mean is not always the best estimate for the KCRV, its properties should be compared with those of alternative estimators. An unjustified application of the weighed mean can result in (1) the attraction of the KCRV to the result having the smallest associated uncertainty, (2) the underestimation of the uncertainty associated with the KCRV and, consequently, (3) the underestimation of the uncertainty associated with the deviation of the result from the KCRV. The last item can lead to an invalid conclusion about the consistency of KC data.

Various approaches can be proposed for KC data evaluation when not all conditions 1–3 are satisfied:

- *Adjustment of the data to the model*. For example, the uncertainties that seem to be unreasonably small can be enlarged or data considered inconsistent with the remainder can be disregarded. Adjustment is reasonable if there is adequate opportunity for thorough analyses of the reasons for uncertainty underestimation and all participants accept those reasons.
- *Generalization of the model* implied by conditions 1–3. Other estimators for the KCRV can be obtained if the correlation between data from different NMIs can be taken into account or other forms of probability distributions can be considered. For this approach the main problem is quantifying the covariance matrix.
- *Use of a different model*. Such a model may require less prior information than implied by condition 1, may not require the mutual independence assumption of condition 2, but would employ condition 3, which is often expected to be reasonable in practice.

This paper addresses the third approach, assuming that:

1. The data from different NMIs are mutually dependent.
2. Some assigned uncertainties can be regarded as invalid.
3. The characteristics of the reproducibility of results inside each NMI are carefully estimated.

A model relating to these conditions is introduced. The interpretation of the KCRV and the DoE are discussed within the context of this model. A solution for the above three problems of KC data evaluation is given.

3. DATA MODEL

A KC is an example of an interlaboratory experiment. Accordingly, the measurement results for each participating NMI can be described by the model given in ISO 5725 [3]:

$$x_i = a + m_i + e_i \ , \tag{3}$$

where x_i, a, m_i and ε_i are, respectively, the measurement result, measurand value, systematic bias and random error in x_i for the ith of the N NMIs participating in the KC.

According to the MRA, “The degree of equivalence … is taken to mean the degree to which these standards are consistent with reference values determined from the KC and hence are consistent with one another”. The consistency with one another means the closeness of the NMIs' results.

The measurand realized by the dispersion of the data for the ith NMI can be described by a distribution function $F_i(x)$, with expectation $a + m_i$ and variance s_i^2. Moreover, the mixture distribution [4]

$$F(x) = (1/N)\sum F_j(x) \tag{4}$$

can be used to describe the distribution for the measurand of which the data provided by the NMIs are realizations. It is not to be confused with a distribution for the measurand relating to the KCRV [2].

Letting $\bar{m} = (1/N)\sum m_i$, the expectation $EF(x)$ and variance of $F(x)$ are, respectively,

$$m = a + \bar{m}, \quad s^2 = (1/N)\sum s_i^2 + (1/N)\sum (m_i - \bar{m})^2. \tag{5}$$

It is suggested that the KCRV is determined as $EF(x)$:

$$x_{ref} = a + \bar{m}. \tag{6}$$

It can be considered as an indicator of the SI value [1]: “… there may be difficulty in relating results to the SI. Nevertheless, the key comparison reference value and deviations from it are good indicators of the SI value”. In the approach here the DoE for the ith NMI,

$$d_i = EX_i - EX = a + m_i - a - \bar{m} = m_i - \bar{m}, \tag{7}$$

can be interpreted as a difference between the “laboratory reference value” and the KCRV, which characterizes the systematic bias of the results of that NMI from the KCRV.

4. DATA ANALYSIS

In the approach here the unbiased estimate of the KCRV is the simple mean $\hat{x}_{ref} = (1/N)\sum x_i$ and $u(\hat{x}_{ref})$ is given by

$$u^2(\hat{x}_{ref}) = u^2((1/N)\sum x_i) = (1/N^2)\sum u_i^2, \qquad (8)$$

where u_i denotes the reproducibility standard uncertainty s_i for NMI i. Accordingly, the DoE and associated uncertainty are given by

$$d_i = x_i - \bar{x},\ \ u^2(d_i) = (1-2/N)u_i^2 + (1/N^2)\sum u_j^2. \qquad (9)$$

In this case the associated uncertainty is caused only by the random dispersions, with standard uncertainty u_i, of the NMI's data. Since it is used for internal quality control, the precision of measurement results inside each NMI is usually characterized sufficiently well. The mutual dependencies between the data due to similar systematic biases are revealed automatically in the form of a mixture distribution.

Even if the NMIs' distributions $F_i(x)$ are Gaussian, $F(x)$ can have a complex form, perhaps asymmetric or multimodal. But, if all the $F_i(x)$ are Gaussian, the distributions for the KCRV and the d_i are also Gaussian.

Thus, to implement the approach, each NMI presents its measurement result and the standard uncertainty due to the associated random dispersion (reproducibility standard deviation). No quantification of covariance is required. The (combined) standard uncertainty associated with the measurement result is not used directly, but is required for confirmation of the stated capabilities of the NMIs.

An advantage of this approach is in preliminary comparisons where there might be doubt about the validity of the provided uncertainties, the comparisons being mainly directed at revealing systematic biases in the NMIs' results. These biases are the predominant reason for the dispersion of results from the NMIs. As in the conventional approach and as stated in section 3, the KCRV can be considered as an estimate of the SI value. If a valid estimate of the SI value is available, estimates of these biases in the NMIs' measurements are obtainable immediately, as are estimates of the systematic deviations between NMIs. The estimation of the SI value is a more general task than that of determining systematic differences between the results from the NMIs. But to address this task requires additional prior information, which, if absent, should not be "invented" to permit solution of the general task. Doing so would result in invalid conclusions about the equivalence of measurement standards and the NMIs' capabilities.

Also, in the context of the given approach the strict solution to the problem of checking the conformity of the KC data with the assigned uncertainties is not generally possible. That solution would again require the covariance matrix of the data to be quantified. However, most of the results would be expected to satisfy the condition

$$\left|x_i - x_j\right| \le 2(u_c^2(x_i) + u_c^2(x_j))^{1/2}, \qquad (10)$$

where $u_c(x_i)$ is the combined standard uncertainty associated with x_i (i.e., including the bias uncertainty). If this check fails in some instances, the results that are inconsistent with the majority of the others should be identified. Then the check of capabilities should be repeated without this result.

5. LINKING OF RMO AND CIPM KEY COMPARISONS

According to the MRA, "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 NMIs in both CIPM and RMO key comparisons". Number the "link NMIs", i.e., those involved in both comparisons, 1 to m. Denote by $x_k^{(1)}$ (= x_k) and $x_k^{(2)}$ the value provided by the kth of these NMIs in the CIPM and RMO comparison, respectively. Then

$$x^{(r)} = (1/m)\sum x_k{}^{(r)}, \quad r = 1, 2, \qquad (11)$$

denote the means in the two comparisons of the values provided by the link laboratories.

The following procedure is suggested:

1. For each NMI participating in both comparisons check the consistency of its realization of the comparisons:
$$\left|x_k^{(2)} - x_k^{(1)}\right| \le 2\sqrt{2}u_k, \quad k = 1, \ldots, m. \qquad (12)$$
2. If the stability check (12) is satisfied, form the DoE ($d_i^{(2)}$, $2u(d_i^{(2)})$) for each NMI that participates only in the RMO comparison from
$$d_i^{(2)} = x_i^{(2)} - x_{ref}, \quad u^2(d_i^{(2)}) = u^2(x_{ref}) + u^2(x_i^{(2)}). \qquad (13)$$
3. Otherwise, form ($d_i^{(2)}$, $2u(d_i^{(2)})$) from
$$d_i^{(2)} = x_i^{(2)} - (x^{(2)} - x^{(1)}) - x_{ref},$$
$$u^2(d_i^{(2)}) = u^2(x_{ref}) + u^2(x_i^{(2)}) \qquad (14)$$
$$+ u^2(x^{(2)} - x^{(1)}) - 2\mathrm{cov}(x^{(1)}, x_{ref}).$$

$u(d_i^{(2)})$ can be derived from (14) in terms of the stipulated values and uncertainties and readily evaluated:

$$u^2(d_i^{(2)}) = u^2(x_i^{(2)}) + \frac{1}{N}\sum_{j=1}^{N} u^2(x_j) + \frac{2(N-m)}{m^2 N}\sum_{k=1}^{m} u^2(x_k). \qquad (15)$$

It is emphasized that the above uncertainties are related only to the random dispersion of the data.

For the laboratories participating in both comparisons, a further estimate of the DoE is available. Its manner of use depends on the purposes of the common participants. One reasonable approach would be to compare the DoEs between the ith and jth NMIs participating in both comparisons, viz., $d_{i,j}^{(1)} = x_i^{(1)} - x_j^{(1)}$ and $d_{i,j}^{(2)} = x_i^{(2)} - x_j^{(2)}$.

6. ALTERNATIVE APPROACHES

There is a further use of a mixture distribution for the evaluation of DoEs, based on the principle of propagation of distributions [5], a generalization of the law of propagation of uncertainty described in the GUM. The application of this principle is reasonable when the distributions of the quantities concerned are not all Gaussian. This case arises when regarding the pooled data as a sample from a mixture distribution, which has a complex form even when the input quantities are Gaussian.

The MRA permits the DoE to be expressed without using a KCRV: "In some exceptional cases, a Consultative Committee may conclude that for technical reasons a reference value for a particular key comparison is not appropriate; the results are then expressed directly in terms of the degrees of equivalence between pairs of standards". This statement implies the estimation of deviations between the NMIs' results. These deviations and the associated uncertainties allow indirectly the prediction of the possible dispersion of results from these NMIs. However, the most general characteristic of the dispersion of such deviations is the probability distribution of the difference. This distribution is a convolution of two distributions, $f_i(x)$ and $f_j(x)$, corresponding to the results from these NMIs: $f(z) = \int f_i(z-x) f_j(x) dx$. It can be useful to consider a tolerance interval as a characteristic of the closeness of results and, thus, as a measure of equivalence. The tolerance interval for the probability distribution $F(x)$ is determined as an interval (Δ_1, Δ_2), which with given probability b encompasses a large fraction p of this distribution $P\left\{\int_{\Delta_1}^{\Delta_2} f(x)dx \geq p\right\} = b$, where $f(x)$ is the probability density corresponding to $F(x)$. There is no principal difficulty in calculating a tolerance interval for a known probability distribution. For example, Monte Carlo simulation can be used. This approach to a quantitative expression of the DoE between two standards can be extended to the case of the group of standards.

Generally speaking, various interpretations of the equivalence of the group of standards are possible:

1. *Overall equivalence - the equivalence of the whole group of the standards*. The mixture distribution $F(x) = (1/N)\sum F_i(x)$ fully characterizes the closeness of the pooled data from the NMIs. A tolerance interval for the mixture presents a partial characterization of the dispersion of the NMIs' results. But the limits of possible deviations of results from different NMIs are even more valuable from a practical viewpoint. A tolerance interval for the difference $X - Y$ of the independent random quantities, each having the distribution $F(x)$, can be used. So a joint characteristic for the whole group is suggested.

2. *Equivalence of each standard to the group*. The equivalence of any one standard to the group of standards can be interpreted as the consistency of results obtained in the NMI with those of the group of NMIs. The DoE can also be expressed by a tolerance interval for the distribution of the difference $X_i - X$ of the result from a particular NMI and that from any NMI in the group, where $X_i \in F_i(x)$ and $X \in F(x)$. In this case every standard has its own DoE with the group of participating standards.

7. CONCLUSIONS

A simplified procedure for evaluating KC data is suggested. It can be used where

1. *There is limited a priori information for quantifying the covariance matrix of the data.* The uncertainties associated with the random reproducibility effects of the NMIs are used rather than the (combined) uncertainties associated with the NMIs' measurements. The approach is indifferent to the various procedures for uncertainty evaluation that may be used by the NMIs. It can be useful, for example, in preliminary comparisons when an uncertainty budget is not agreed by all NMIs.

2. *There is no clear interpretation (understanding) of the reference value from the physical point of view.* The KCRV is interpreted as a conventional value for the KC, close to a case where it solely provides an uncertainty-free datum for presenting the KC results. In the given approach only the variability of the KRCV due to random effects in the NMIs is taken into account.

3. *Equivalence can be interpreted as a certain level of reproducibility of results obtained in the particular group of NMIs.* An estimate is obtained, following the approach here, of the reproducibility of results from the NMIs. For each NMI it is expressed as a systematic deviation from the KCRV and the uncertainty of this deviation. Also, another equivalence characteristic can be proposed using a mixture distribution. The advantage of a tolerance interval as a measure of reproducibility of results is emphasized: it provides directly limits within which with a high probability a significant number of the measurement results obtained by the NMIs (or of the deviations between results from particular NMIs) lie.

A model-based approach based on related concepts has also been considered in the context of the pair-wise comparison of NMIs [6].

REFERENCES

[1] BIPM, "Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes", Bureau International des Poids et Mesures, Sèvres, France, 1999.

[2] M. G. Cox, "The evaluation of key comparison data", *Metrologia*, vol. 39, pp. 589–595, 2002.

[3] ISO, "ISO 5725, Accuracy (trueness and precision) of measurement methods and results", International Organisation for Standardisation, Geneva, Switzerland, 1994.

[4] A. Chunovkina and A. Chursin, "GUM and MRA: Some problems of data processing and measurement uncertainty evaluation", in "Advanced Mathematical Tools in Metrology V. Series on Advances in Mathematics for Applied Sciences", vol. 57 (edited by P. Ciarlini, M. G. Cox, E. Felipe, F. Pavese and D. Richter), pp. 55–62, World Scientific, Singapore, 2001.

[5] JCGM, Guide to the Expression of Uncertainty in Measurement. Supplement 1. Numerical Methods for the Propagation of Probability Distributions. Joint Committee for Guides in Metrology, 2002. Draft

[6] D.R. White, "On the choice of comparison reference values for the purpose of pair-wise comparison of laboratories", CPEM 2000, Conference on Precision Electromagnetic Measurements, Sydney, May 2000 (edited by J. Hunter and L. Johnson), pp. 325-326, IEEE, 2000.

AUTHORS:
Anna Chunovkina
Head of the sector "Statistical methods in metrology"
D.I. Mendeleyev Institute for Metrology (VNIIM),
Moskovsky pr. 19, St.Petersburg, 198005, Russia
Tel: +7(812)3150984 Fax: +7(812)1130114,
A.G.Chunovkina@vniim.ru
Maurice Cox
Principal Science Leader in Mathematics and Statistics
National Physical Laboratory,
Teddington, Middlesex TW11 0LW, UK
Tel: +44(20)8943 6096, Fax: +44(20)8977 7091,
maurice.cox@npl.co.uk