APEM journal
Advances in Production Engineering & Management
Volume 9 | Number 1 | March 2014 | pp 44–52
http://dx.doi.org/10.14743/apem2014.1.175
ISSN 1854-6250
Journal home: apem-journal.org
Original scientific paper

Performance metrics for testing statistical calculations in interlaboratory comparisons

Acko, B.[a,*], **Sluban, B.**[a], **Tasič, T.**[a], **Brezovnik, S.**[b]

[a]Faculty of Mechanical Engineering, University of Maribor, 2000 Maribor, Slovenia
[b]Gorenje gospodinjski aparati, d. d., 3503 Velenje, Slovenia

ABSTRACT

Interlaboratory comparisons are the most powerful tools for determining the competences of laboratories performing calibrations and testing. Performance metrics is based on statistical analysis, which can be very complex in certain cases, especially for testing where transfer standards (samples) are prepared by the pilot laboratory. Statistical quantities are calculated using different kinds of software, from simple Excel applications to universal or specific commercial programmes. In order to ensure proper quality of such calculations, it is very important that all computational links are recognized explicitly and known to be operating correctly. In order to introduce a traceability chain into metrology computation, the European project EMRP NEW 06 TraCIM was agreed between the EC and the European Metrology Association (EURAMET). One of the tasks of the project was also to establish random datasets and validation algorithms for verifying software applications in regard to evaluating interlaboratory comparison results. The statistical backgrounds for resolving this task, and the basic concept of the data generator are presented in this paper. Background normative documents, calculated statistical parameters, boundary conditions for generating reference data sets are described, as well as customer interface.

ARTICLE INFO

Keywords:
Interlaboratory comparisons
Data generator
Software validation

**Corresponding author:*
bojan.acko@um.si
(Acko, B.)

Article history:
Received 5 September 2013
Revised 17 January 2014
Accepted 10 February 2014

1. Introduction

Interlaboratory comparisons are used to determine the performance of individual laboratories for specific calibrations, tests or measurements, and to monitor the continuing performance of laboratories [1-3]. In statistical language, the performance of laboratories can be described by three properties: laboratory bias, stability and repeatability [3]. In calibrations, stability and repeatability are normally substituted by the measurement uncertainty, estimated and reported by participating laboratories [1, 2]. In testing, laboratory bias may be assessed by tests on reference materials, when these are available. Otherwise, interlaboratory comparisons provide a generally available means of obtaining information about laboratory bias. However, stability and repeatability will affect data obtained in comparison, so that it is possible for a laboratory to obtain data in a round of a proficiency test which indicate bias that is actually caused by poor stability or poor repeatability [3].

One of the tasks of the European research project EMRP NEW 06 TraCIM [4] is to establish random datasets and validation algorithms for verifying software applications for evaluating interlaboratory comparison results. This task is shared between the Laboratory for Production Measurement at University in Maribor and the German national metrology institute PTB. The

article aims to present general ideas and approaches on establishing an internet-based application, which will serve organizers of different kinds of interlaboratory comparisons to check correctness of their evaluation algorithms based on standardized and other internationally recognized statistical procedures.

2. Statistical quantities in interlaboratory comparisons

2.1 Goals of interlaboratory comparisons

An interlaboratory comparison is a computationally-intensive metrological tool for evaluating performance of different kinds of metrological laboratories, from national metrology institutes to market-oriented calibration and testing laboratories. Approaches in organization and statistical evaluation of the results can be very different and depend on the aim of an interlaboratory comparison, number of participants, their quality, form of results, etc. [5, 6]. Special approaches are used for international key comparisons for evaluating performance quality of national metrology institutes. The application of the procedures to a specific set of key comparison data provides a key comparison reference value (KCRV) and the associated uncertainty, the degree of equivalence of the measurement made by each participating national institute and the degrees of equivalence between measurements made by all pairs of participating institutes [1, 2, 5, 6]. On the other hand, interlaboratory comparisons applied in proficiency testing of testing laboratories follow standardized procedures [3, 7, 8] recommending different statistical evaluations of results for different types of interlaboratory comparison and for different ways of reporting measurement results.

2.2 Performance metrics

In order to evaluate performance of the participants in an interlaboratory comparison, measurement data (with or without associated measurement uncertainties) shall be collected from all the participants and evaluated by means of an agreed statistical approach [1-10]. Single measurement values reported by participants are compared with an agreed assigned (reference) value by considering reported measurement uncertainties and the uncertainty of the assigned value. The basic principle of evaluating performance of participants in an interlaboratory comparison is shown in Fig. 1 [7]. Different cases of reporting measurement results shall be considered. In BIPM key or supplementary comparisons and other calibration comparisons, one result and the assigned measurement uncertainty are reported [1, 2, 9, 10], while in some cases of comparisons in testing participants report more results without uncertainty. The assigned (reference) value can be calculated from the reported measurement values or simply defined as a value of the reference material or as a measurement value of the reference laboratory. The performance metrics depends on the way of reporting results and defining the assigned value. The uncertainty of the assigned value shall be considered in all cases. This value might be substituted by a standard deviation of the intercomparison scheme [3]. Uncertainties or other forms of dispersions of reported results shall be considered as well. These values are declaring quality of performance of participating laboratories in the scheme. In most cases participating laboratories declare their quality by themselves by reporting standard or expanded uncertainty or by reporting more results of the same measurand. However, in some cases in testing area pilot laboratory defines allowed deviation of reported results from the assigned value. In such cases participants don't report uncertainty of measurement [3].

Interlaboratory comparisons are statistically evaluated by using diverse software, which might produce errors in final results. Error sources could be computational malfunctions, typing mistakes, mistakes in statistical formulae, etc. In order to detect such errors, reference data sets and algorithms for all possible statistical approaches should be produced and made available to the pilots of interlaboratory comparisons, who are responsible to perform reliable performance metrics. Such reference data sets and calculations shall be cross-checked by using different software packages and by comparisons in different institutes [4].

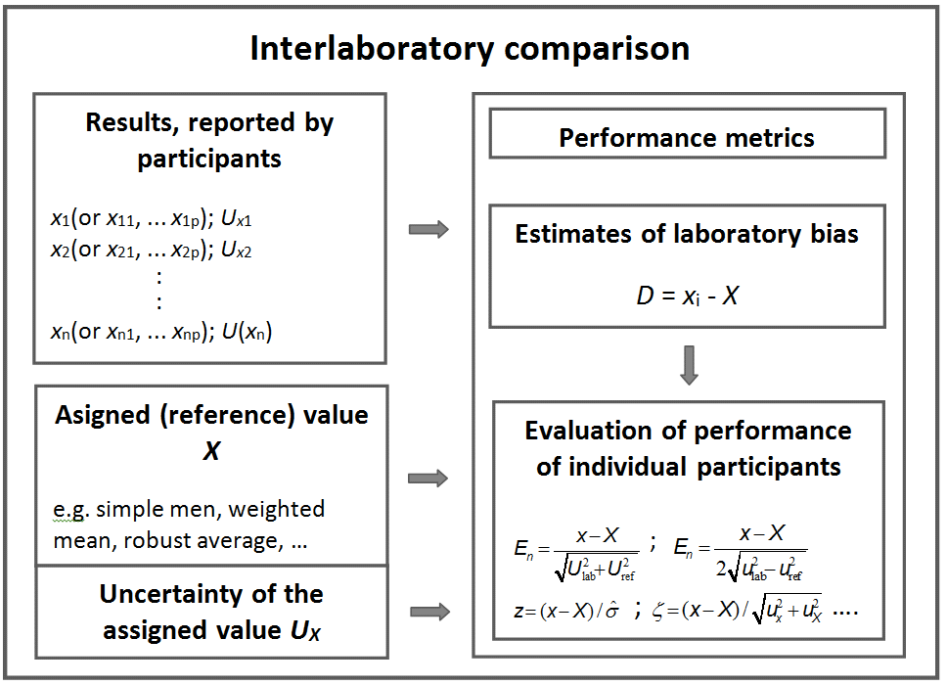

Fig. 1 Evaluation of interlaboratory comparison results

2.3 Reported values

Input values into the evaluating software are measurement results reported by participants and corresponding measurement uncertainties, as well as boundary conditions and evaluation strategy. Participants can report one or more results for the measurement quantity [2, 3, 7]:

- $x_1, x_2, \ldots, x_n$

  or

- $x_{11}, x_{12}, \ldots, x_{1n}$
  $x_{21}, x_{22}, \ldots, x_{2n}$
  $\vdots$
  $x_{p1}, x_{n2}, \ldots, x_{pn}$

The results can be reported with or without measurement uncertainties. The uncertainties can be reported as standard uncertainties ($u_{x1}, u_{x2}, \ldots, u_{xn}$) or expanded uncertainties at certain level of confidence ($U_{x1}, U_{x2}, \ldots, U_{xn}$) [11]. Uncertainties are not reported in some cases of comparisons in testing, especially when more than one result is reported per participant [3, 7].

Output values are calculated in accordance with the users' needs. User can select the set of output values through an intelligent interface. Generated input values are also considering user's boundary conditions. Most common output values in accordance with international standards and recommendations are presented in the following chapters [7].

2.4 Assigned value

The way of determining an assigned (reference) value should be defined prior to the interlaboratory comparison. The value can be determined in advance as a "certified reference value" X_{CRM} (when the material used in a proficiency test is a certified reference material) [3, 7] or a "reference value" X_{RM} (a value of the prepared reference material derived from a calibration against the certified reference values of the CRMs) [3, 7] or a "consensus value from expert laboratories" [3]. However, the most common way of determining the assigned value in calibration interlaboratory comparisons is to calculate it from the reported results x_i as a simple mean [1, 2, 9]:

$$X = \bar{x} \tag{1}$$

or a weighted simple mean [1, 2, 9]:

$$X = \frac{\sum_{i=1}^{p} u^{-2}(x_i) \cdot x_i}{\sum_{i=1}^{p} (x_i)} \tag{2}$$

where:

x_i – measured results reported by participants

$u(x_i)$ – uncertainties of the measured results reported by participants

When the participating laboratories report more than one measured value without stating uncertainty of measurement (in proficiency tests of testing laboratories), the reference value is calculated as a "robust" average [3, 7]:

$$X = \sum x_i^* / p \tag{3}$$

where:

$$x_i^* = \begin{cases} x^* - \delta, & \text{if } x_i < x^* - \delta \\ x^* + \delta, & \text{if } x_i < x^* + \delta \\ x_i \quad , & \text{otherwise} \end{cases}$$

x^* – median of x_i (i = 1, 2,..., p)

δ = 1.5s^*

$x_1, x_2, \ldots, x_p$ – items of data, sorted in ascending order

2.5 Uncertainty of the assigned value

The assigned value is always determined experimentally with certain uncertainty. The uncertainty depends on the way of determining the assigned value and is calculated by following standardized or internationally recognized procedures [1, 2]. If the assigned value is calculated as a simple mean, its standard uncertainty is [1, 2, 9]:

$$u(X) = \sqrt{\frac{\sum u^2(x_i)}{p}} \tag{4}$$

where:

$u(x_i)$ – uncertainties of the measured results reported by participants

p – number of participants

Standard uncertainty of the weighted mean is [1, 2, 9]:

$$u(X) = \frac{1}{\sqrt{\sum_{i=1}^{n} u^{-2}(x_i)}} \tag{5}$$

where:

$u(x_i)$ – uncertainties of the measured results reported by participants

p – number of participants

The robust average has the following standard uncertainty assigned to it [3, 7]:

$$u(X) = 1.25 \cdot \frac{s^*}{\sqrt{p}} \; ; \; u(x_i) \text{ not reported} \tag{6}$$

$$u(X) = \frac{1.25}{p} \sqrt{\sum_{i=1}^{p} u(x_i)^2} \; ; \; u(x_i) \text{ reported} \tag{7}$$

where:

$$x_i^* = \begin{Bmatrix} x^* - \delta, & if\ x_i < x^* - \delta \\ x^* + \delta, & if\ x_i < x^* + \delta \\ x_i \quad , & otherwise \end{Bmatrix}$$

x^* – median of x_i $(i = 1, 2, \ldots, p)$

$\delta = 1.5s^*$

$x_1, x_2, \ldots, x_p$ – items of data, sorted into increasing order

$$s^* = 1.134\sqrt{\sum (x_i^* - x^*)/(p-1)}$$

If the assigned value is defined by perception, the following standard deviation is assigned to it [3, 7]:

$$\hat{\sigma} = \sqrt{(\emptyset \times \sigma_L)^2 + (\sigma_r^2/n)} \tag{8}$$

where:

$\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2}$

σ_R – reproducibility standard deviation

σ_r – repeatability standard deviation

n – number of replicate measurements each laboratory is to perform

$\sigma_R = 0{,}02c^{0.8495}$

c – concentration of chemical species to be determined in percent (mass fraction)

In the case of defining the assigned value from the results of a precision experiment, the standard deviation is expressed as [3, 7]:

$$\hat{\sigma} = \sqrt{\sigma_L^2 + (\sigma_r^2/n)} \tag{9}$$

where:

$\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2}$

σ_R – reproducibility standard deviation

σ_r – repeatability standard deviation

n – number of replicate measurements each laboratory is to perform

2.6 Performance statistics

Performance statistics is used for evaluating performance of participating laboratories. The final result for single laboratory is most usually "passed" or "failed". Corrective actions shall be taken by the laboratory, which fails the interlaboratory comparison. The first step in the performance statistics is to evaluate estimates of laboratory bias. An estimate could be evaluated as an absolute difference [3, 6, 7]:

$$D = x - X \tag{10}$$

where:

x – result reported by a participant

X – assigned value

or as a percentage difference [3, 7]:

$$D_{\%} = 100 \cdot (x - X)/X \tag{11}$$

The laboratory bias is tan used in different types of evaluation parameters, which should be in certainty limits in order to pass the comparison.

z-score is used in proficiency testing, where no uncertainties are reported by participating laboratories. The z-score is calculated by the following equation [3, 7]:

$$z = (x - X)/\hat{\sigma} \tag{12}$$

where:

x – result reported by a participant

X – assigned value

$\hat{\sigma}$ – standard deviation for proficiency assessment

When a participant reports a result that gives rise to a z-score above 3.0 or below −3.0, then the result shall be considered to give an "action signal". Likewise, a z-score above 2.0 or below −2.0 shall be considered to give a "warning signal". A single "action signal", or "warning signals" in two successive rounds, shall be taken as evidence that an anomaly has occurred that requires investigation.

E_n numbers are used in the comparisons, in which participating laboratories report measurement uncertainties in accordance with the Guide to the expression of uncertainty in measurement (GUM). If the reference value X is calculated as a simple mean, the following equation is used [1, 2, 9, 10]:

$$E_n = \frac{x - X}{\sqrt{U_{lab}^2 + U_{ref}^2}} \tag{13}$$

where:

U_{ref} – expanded uncertainty of the reference value X

U_{lab} – expanded uncertainty of a participant's result x

If the reference value X is calculated as a weighted mean, the E_n value is calculated as follows [1, 2, 9, 10]:

$$E_n = \frac{x - X}{2 \cdot \sqrt{u_{lab}^2 - u_{ref}^2}} \tag{14}$$

where:

u_{ref} – standard uncertainty of the reference value X

u_{lab} – standard uncertainty of a participant's result x

When uncertainties are estimated in a way consistent with the Guide to the expression of uncertainty in measurement (GUM), E_n numbers express the validity of the expanded uncertainty estimate associated with each result. A value of $|E_n| < 1$ provides objective evidence that the estimate of uncertainty is consistent with the definition of expanded uncertainty given in the GUM.

z'-scores are used when the assigned value is not calculated using the results reported by the participants and when the participants don't report uncertainties of their results. The z'-score is calculated by the following equation [3, 7]:

$$z' = (x - X)/\sqrt{\hat{\sigma}^2 + u(X)^2} \tag{15}$$

where:

x – result reported by a participant

X – assigned value

$\hat{\sigma}$ – standard deviation for proficiency assessment

$u(X)$ – standard uncertainty of the assigned value X

z'-scores shall be interpreted in the same way as z-scores using the same critical values of 2.0 and 3.0.

ζ-scores are used when the assigned value is not calculated using the results reported by the participants and when the participants report uncertainties of their results. The ζ-score is calculated by the following equation [3, 7]:

$$\zeta = (x - X)/\sqrt{u(x)^2 + u(X)^2} \tag{16}$$

where:

$u(x)$ – laboratory own estimate of the standard uncertainty of the result x

$u(X)$ – standard uncertainty of the assigned value X

When there is an effective system in operation for validating laboratories' own estimates of the standard uncertainties of their results, ζ-scores may be used instead of z-scores, and shall be interpreted in the same way as z-scores, using the same critical values of 2.0 and 3.0.

Another criteria are E_z-score [3]. Both values E_{z-} and E_{z+} shall be between -1 and 1 in order to be able to claim that the participating laboratory performance is satisfactory.

$$E_{z-} = \frac{x - (X - U(x))}{U(X)} \quad \text{and} \quad E_{z+} = \frac{x - (X + U(x))}{U(X)} \tag{17}$$

where:

x – result reported by a participant

X – assigned value

$U(x)$ – expanded uncertainty of the result x

$U(X)$ – expanded uncertainty of the assigned value X

2.7 Additional parameters

Additional parameters to be evaluated by the interlaboratory comparison evaluation software are different kinds of significance tests (χ^2-test, Birge criterion, etc.), confidence ellipse, rank correlation test, repeatability standard deviations [3, 9, 10].

3. Software validation application

The interlaboratory comparison software validation application allows the user to define boundary conditions for generating random data sets, for which selected reference statistical quantities are calculated. The user's software is then validated by comparing reference quantities with those calculated by the user's software [7].

The software validation application was developed in the environment Microsoft Visual Studio.Net 2012. Rounding of calculated data is defined based on the data type, declaration and other data properties. Uncertainty of calculated data will be defined by comparing calculation results with results created in other environments (e.g., Mathematica) and with results gained by other project partners.

3.1 User's interface

The user's interface consists of three modules [7]:

- selection of boundary conditions for generating data sets,
- selection of statistical quantities to be calculated,
- computation of statistical quantities and graphical presentation.

The first module is allowing the customer to define the following interlaboratory comparison characteristics [7]:

- number of participants, which is not limited,
- information about reporting uncertainty of measurement,
- number of results reported by single participant (this value is automatically set to 1, if "yes" is selected in the previous row),
- target value for the reported result (normally nominal value of the measurand or predefined assigned value),
- variation of results (this value can be extracted from real intercomparison results),
- accuracy of results (number of decimal places is selected based on the knowledge about real interlaboratory comparison results),
- type of measurement uncertainty (standard or expanded; the coverage factor can be selected in the case of expanded uncertainty),
- variation of the measurement uncertainty (this value can be extracted from real interlaboratory comparison results),
- accuracy of measurement uncertainty (number of decimal places is selected based on the knowledge about real interlaboratory comparison results).

3.2 Data generator

The data generator generates a random set of data after all boundary conditions are defined. This data set contains all numerical characteristics of real interlaboratory comparison results. Generation of random data sets can be repeated unlimited number of times. Generated data are real numbers with selectable number of decimal places. Uncertainty of generated data is $u = 0$, since the data is not rounded and since single values are independent from each other.

In order to reflect real interlaboratory conditions, generation of data sets is not completely randomly within boundary conditions selected by the user. The data is approximately normally distributed around the selected target value. Furthermore, one or more outliers can be incorporated into the data set.

After the data set is generated, the user can select statistical quantities (Section 2) to be verified. In the final module, the customer can see reference results and their graphical presentation [7].

4. Conclusion

The application consisting of data set generator and calculator of statistical quantities for interlaboratory comparisons is still being developed in the frame of running EMRP TraCIM project. The first module for selecting boundary conditions for creating data sets has already been finished and agreed among the participants in the corresponding project work package. Some modifications still need to done in the data generator. Generation of normally distributed values with incorporated outliers is still under consideration. After finishing the second and the third module, the calculation results will be validated by means of using different kinds of software and by comparison among the project participants.

The universal on-line application for validating different software packages for interlaboratory comparison data calculation is planned to be a free accessible internet application, which will be aimed to serve organizers of all interlaboratory comparisons, who are following standardized or internationally recognized rules. The main purpose of using the presented application will be to avoid misinterpretations of interlaboratory comparison results that might lead to wrong evaluation of the participants' performance capability. Therefore, the application will help to improve international comparability and traceability of measurement results in all types of proficiency testing.

Acknowledgement

The authors would like to acknowledge funding of the presented research within the European Metrology Research Programme (EMRP) in the Joint Research Project NEW06 TraCIM, as well as funding of national research in the frame of the national standard of length by the National Metrology Institute of Republic of Slovenia (MIRS). Furthermore, fruitful professional discussions within the research group, especially with Phisikalisch Technische Bundesanstalt, are highly appreciated.

References

[1] Cox, M.G. (2002). The evaluation of key comparison data, *Metrologia,* Vol. 39, No. 6, 589-595, doi: 10.1088/0026-1394/39/6/10.

[2] BIPM (2014). Guide for implementation of the CIPM MRA, CIPM MRA-G-01, v. 1.1, from *http://www.bipm.org/en/cipm-mra/documents*, accessed January 10, 2014.

[3] ISO 13528 (2005). *Statistical methods for use in proficiency testing by interlaboratory comparisons*, ISO copyright office, Geneva.

[4] Forbes, A. (2012). *NEW06 TraCIM – Traceability of computationally-intensive metrology, EMRP JRP Protocol*, NPL, Teddington.

[5] Cox, M.G. (2007). The evaluation of key comparison data: determining the largest consistent subset, *Metrologia*, Vol. 44 , No. 3, 187-200, doi: 10.1088/0026-1394/44/3/005.

[6] Cox, M.G., Harris, P.M. (2012). The evaluation of key comparison data using key comparison reference curves, *Metrologia*, Vol. 49, No. 4, 437-445, doi:10.1088/0026-1394/49/4/437.

[7] Acko, B., Brezovnik, S., Sluban, B. (2013). Verification of software applications for evaluating interlaboratory comparison results, In: *Annals of DAAAM International for 2013, Collection of working papers for 24th DAAAM International Symposium*, DAAAM International Vienna, Vienna, 9 pages.

[8] ISO 5725-1 (2001). *Accuracy (trueness and precision) of measurement methods and results – Intermediate measures of the precision of a standard measurement method*, ISO copyright office, Geneva.

[9] Härtig, F., Kniel, K. (2013). Critical observations on rules for comparing measurement results for key comparisons, *Measurement,* Vol. 46, No. 9, 3715-3719, doi: 10.1016/j.measurement.2013.04.079.

[10] Acko, B. (2012). Final report on EUROMET key comparison EUROMET L-K7: calibration of line scales, *Metrologia*, Vol. 49, No. 1A, Technical Supplement, doi: 10.1088/0026-1394/49/1A/04006.

[11] Raczynski, S. (2011). Uncertainty, dualism and inverse reachable sets, *International Journal of Simulation Modelling*, Vol. 10, No. 1, 38-45, doi: 10.2507/IJSIMM10(1)4.180.