XVIII IMEKO WORLD CONGRESS
Metrology for a Sustainable Development
September, 17 – 22, 2006, Rio de Janeiro, Brazil

SOME NOTES ON REPLICATED MEASUREMENTS IN METROLOGY AND TESTING: CLASSIFICATION INTO REPEATED OR NON-REPEATED MEASUREMENTS

Franco Pavese

Istituto Nazionale di Ricerca Metrologica (INRIM)[a], Torino, Italy, f.pavese@inrim.cnr.it

Replication of measurements and the combination of observations are standard and essential practices in metrology. They are done with different methods to match distinct purposes:

(a) When done on the same standard, to obtain a statistics allowing to assess the repeatability of the value of standard;

(b) When done on the same standard, to check for the influence on the total uncertainty arising from the variability of the influence parameters affecting the standard, including dependence on time. The checks allow recording the day-to-day standards history;

(c) When done on several standards of the same Laboratory, to check if they have the same value or to establish the differences between their values, and evaluate the associated uncertainty.

Exercise (c) can be called *intra*-laboratory comparison.

When the same exercise is performed for directly comparing one (or more) standards provided by different laboratories, it is called *inter*-laboratory comparison.

When the exercise is performed to assess “periodically the overall performance of a laboratory” [1], i.e. that the Laboratory can continuously demonstrate its ability to correctly conduct a certain type of measurement, the exercise should be considered a proficiency test, a usual exercise in the testing field.

The GUM [2] is defining “repeated measurements” by saying that the uncertainty components in category A are “those which are evaluated by applying statistical methods to a series of repeated determinations”.

This definition is equivalent to say that the repeatability of replicated measurements only includes components of category A of the total uncertainty budget.

In fact, according to the last draft available of the VIM [3], “Type A evaluation of measurement uncertainty” arises from “a statistical analysis of the quantity values obtained by measurements under repeatability conditions” (2.13).

According to ISO 3534-1 as also reported in ISO 5725-1:1994 [4], “repeatability conditions” are “conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time”. This definition is also adopted by VIM (2.36).

In other words, the values of all the influence factors/parameters are assumed not to change during those intervals of time. This situation has also been indicated as the fact that all the measurements can be considered to be obtained at the same “experimental unit” [5].

When the J replicated measurements made by a i-th laboratory are under these conditions, one can combine them by using a data model. It is different for the metrology and the testing frames.

In calibration, it is written:

$$y_{ji} = a + \varepsilon_{ji} \qquad j = 1 \ldots J \tag{1}$$

where y, the output estimate of the measurand value, is a $f(x_1 \ldots x_n)$, being x_i the “measurable quantities” [1], a is the measurand value –always unknown and not measurable by definition– and ε_j is the zero-mean random error occurring at every j measurement. The replication of the observations allows to gain a knowledge of the statistics of Y, and, by increasing the number of replications, the estimated standard error can be reduced.

In testing [3], it is written:

$$y_{ji} = m + B_i + \varepsilon_{ji} \qquad j = 1 \ldots J \tag{2}$$

where “m is the general mean (expectation); B is the laboratory component of bias under repeatability conditions; ε is the random error occurring under repeatability conditions”.

The two models are, in principle, not conflicting by noting that GUM indicates that “it is assumed that each input estimate x_i is corrected for all known systematic effects that significantly influence the estimate y”, i.e. all $B_i = 0$, the assumptions are different.

However, case (a) does not allow one to tell anything about reproducibility and accuracy. In particular, (c) is the case when one is examining (*intra*- or *inter*-) comparison data. Therefore cases (b) and (c) are the essence of metrology and testing data assessment. In metrology, (c) assumed a peculiar importance in the definition of the “degree of equivalence” between laboratories with the “key (*inter*-)comparisons” set up in the frame of the MRA [6]. A review of the problems arising from the needs prompted by MRA can be found in [7].

[a] Until 31.12.2005, Istituto di Metrologia “G.Colonnetti” (IMGC), CNR.

The replication of measurements outside case (a) involves “Type B evaluation of measurement uncertainty”, defined by VIM “method of evaluation … by means other than a statistical analysis of quantity values obtained by measurement” (2.14) [3]. GUM definition (2.3.3) is similar. This definition, which is clearly opposite to considering repeated measurements, seems to involve only an expert judgment. It seems to exclude measurement reproducibility, case (b), defined by VIM as “measurement precision under reproducibility conditions of measurements” (2.41) that are those “in a set of conditions using different locations, operators and measuring systems”, the latter including the “use of different measurement procedures” (2.40). For testing, ISO 5725-1 (3.18) states the same, except the latter condition. A statistical analysis is obviously necessary also on reproducibility data. However, GUM states that “Type A evaluations of standard uncertainty components are founded on frequency distributions, while Type B evaluations are founded on *a priori* distributions”, both being “models that are used to represent the state of the knowledge” (4.1.6). The non-repeated measurements performed to evaluate the reproducibility do not seem to fall into neither categories.

The comparison of replicated measurements according to case (c) is aimed instead at evaluating the accuracy of measurement. The latter is defined by VIM as “closeness of agreement between a quantity value obtained by measurement and the true value” (3.5), the true value of a quantity being “quantity value consistent with the definition of a quantity” (1.19) and noting that “within the Classical Approach a unique value is thought to describe the measurand … is by nature unobtainable” and that “due to definitional measurement uncertainty, there is a distribution of true values consistent with the definition of a measurand … by nature unknowable”. In testing, very often there is an intrinsic differencein that, the “true value” can be assigned to the measurand: in fact, ISO 5725-1 defines a trueness (3.7) in an operational way as “the closeness of agreement between the average value obtained from a large series of test results and an accepted reference value”, ISO 5725-5 is dedicated to this issue and ISO 5725-4 is integrating model (2) by the following:

$$y = \mu + \delta + B + \varepsilon \qquad (3)$$

where now “μ is the accepted reference value of the property being measured, δ is the bias of the measurement method” and the rest is as in model (2). Here μ is equivalent to a in model (1) but is supposed to be known.

In conclusion, it seems that, for an exercise of type (b) performed on a sample/device (a standard in metrology), a model similar to (1) can be written as:

$$y_{ji} = a + \varepsilon_{ji} + \eta_{ji} \qquad j = 1 \ldots J \qquad (4)$$

where ε is the random uncertainty part arising from the repeated measurements and η is the one arising from the additional non-repeated measurements obtained by testing with a suitable procedure (b) –possibly integrated by expert estimate– the effect of the variability of the influence factors. In fact, no procedure suitable to test measurement precision can provide evidence that the evaluation of Type B uncertainty component due the total variability of the influence factors is zero-mean or not.

Different is the case of an exercise of type (c), an (*intra-* or *inter-*) comparison of samples. The aim is here to assess accuracy, i.e., to perform an evaluation of Type B component of uncertainty that includes an evaluation of the differences between the expectation values assigned by the laboratories to their samples/devicces. In fact, should even every laboratory have “corrected for all known systematic effects” its value, these “Type B evaluations are founded on … models that are used to represent the state of the knowledge” [2], knowledge that is in general insufficient within each laboratory to assess accuracy. In other words, the past usual experience makes one have to assume, as an *a priori* knowledge, that the comparisons will in fact show differences in the values of the samples/devices assigned by each laboratory.

This is already taken into account in the model used by ISO 5725 in the testing field –model (2) or (3).

On the contrary, in calibration there are at present two ways of thinking on this issue.

According to the first way, model (1) should apply also to comparisons and then one should test for *consistency* of the data with the assumption of repeated-measurements. Often, a χ^2-test is proposed for this purpose, which underpins other strong assumptions, namely Normality. Notice that exercise (c) includes the Type B of uncertainty evaluation. Test failure would not pass the hypothesis that the measurements can be considered repeated.

According to the second way, the *a priori* knowledge is used based on the most usual evidence that for most of the comparisons “when the *i*-th participant repeats the comparison *j* times, then its results can be distributed about an expectation value differing from the measurand value a by an amount m_i with standard deviation s_i” [8], where m_i has the meaning of $(m + B_i)$ in model (2). In other words, the basic model for a case (c) exercise becomes the following:

$$y_{ji} = = a + m_i + \varepsilon_{ij} \qquad i = 1 \ldots I, j = 1 \ldots J \qquad (5)$$

where the subscript i refers to the *i*-th standard. The term m_i not depending on j (having the same value μ_i for all $j = 1 \ldots J$) is expressing the variability of the *i*-th standard and is telling us that part of the model does not apply to all the measurements, but only to the subset concerning the specific *i*-th standard. In this case, after an estimate of the differences between values μ_i are obtained[1], one should then check for their *compatibility*. The definition of the latter according to VIM is “property satisfied by all the measurement results of the same quantity, characterised by an adequate overlap of their corresponding sets of quantity values” (2.30). Test failure for some $\mu_h - \mu_k$ would indicate failure of *consistency* with the hypothesis of they being significantly different from zero. This is not allowed for the key comparisons according to MRA, where the differences have the meaning of “degrees of equivalence”, a non-hierarchical concept.

[1] Actually, μ_i remain as unknown as a is, only the differences $(\mu_h - \mu_k)$ of pairs of laboratories are measured.

REFERENCES

[1] European Accreditation, EA Guidelines on the expression of uncertainty in quantitative testing, EA-4/16, December 2003 rev00.

[2] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, and OIML, Guide to the Expression of Uncertainty in Measurement, International Standards Institution, Second Edition, 1995.

[3] International Vocabulary of Basic and General Terms in Metrology (VIM), III Edition, draft of April 2004, BIPM, 2004.

[4] ISO 5725, Accuracy (trueness and precision) of measurement methods and results (Geneva: International Organization for Standardization) 1994.

[5] ISO 3534-3 1999 Statistics – Vocabulary and Symbols – Part 3: Design of Experiments 2nd ed (Genève: International Organization for Standardization) pp 5 (1.9), 31 (2.6) and 40–42 (3.4)

[6] CIPM, Mutual Recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, Bureau International des Poids et Mesures, Sèvres, 1999.

[7] F. Pavese, A metrologist viewpoint on some statistical issues concerning the comparison of non-repeated measurement data, namely MRA Key Comparisons, Measurement 2006, Special Issue, in print.

[8] D.R. White, CPEM 2000 14-19 May, Sydney, CPEM conference digest pp. 325-326; Metrologia 41 (2004),122-131.