ADRESSING COMPLEXITY IN KEY COMPARISON ANALYSIS

Jennifer Decker, Alan Steele, and Rob Douglas
National Research Council of Canada, Institute for National Measurement Standards
1200 Montreal Road, Ottawa, CANADA K1A 0R6
Telephone: +1 613 991 1633, Fax: +1 613 952 1394 and e-mail: jennifer.decker@nrc-cnrc.gc.ca

Abstract: The conclusions from international key comparison (KC) experiments constitute an important basis for validating the mutual recognition of measurement capabilities between the national metrology institutes (NMIs). Conventional χ^2 (chi squared) methods for data analysis have been recommended in a move to promote uniformity in the conclusions drawn from KC results. Extended χ^2 methods can address the complexities of degrees of freedom or novel algorithms for using peer results to select a reference value. In many situations, the utility of a KC reference value can be overstated. By focusing instead on pair differences, *unmediated* consistency analysis offers a better method for testing the statistical consistency of a KC data set. This paper discusses extended χ^2 methods, using the mean-square of the normalized differences between pairs of measurements, or pairs of "consensus invariants" derived from the measurements. The underlying simplicity of comparing real measurements to each other (avoiding any equivocal approximation to "the right answer") can reduce the perceived complexity of a KC experiment, particularly for KCs with multiple measurands that can all be incorporated into the χ^2 average. An example from CCL.SIM-K1 is discussed.

1. INTRODUCTION

One of the principal goals of key comparisons undertaken in support of the CIPM mutual recognition arrangement (MRA) [1] is to demonstrate inter-laboratory consistency in the principal techniques used in various major metrology areas. Typically, an artifact is circulated for measurement, and each of the *N* NMIs participating in the key comparison submits the value of their measurement result and an associated uncertainty. For these KCs, chi-squared statistics [2, 3] are a widely appreciated way to measure and communicate the consistency of the *N* results, by comparing the dispersion of the *N* reported values with the dispersion expected from the *N* claimed uncertainties. The implicit metrology conjecture is simple: the measurement values reported by the NMIs are expected to agree within the associated uncertainties they reported.

Dimensional metrology provides concrete examples for how complexity in a KC can be managed. In dimensional metrology, the principal technique for measuring gauge block length with direct traceability to the SI definition of the metre is optical interferometry. This case illustrates clearly the basic consistency test using a chi-squared methodology for analysis of a scalar measurand.

In dimensional metrology there are also a variety of more complex comparisons that include multiple measurements taken simultaneously. They make use of more sophisticated artifacts such as linear scales, or ball plates for co-ordinate metrology. In some cases, the multiple measurements may be presented explicitly as a vector with an intuitive physical meaning (the ball plate, for example). In other cases the multiple measurements may be presented as a column in a table that might also be treated as a vector quantity, but more as a matter of notational convenience than to evoke an intuitively appealing physical appreciation of the measurands.

In dimensional metrology, it is also common practice to circulate multiple artifacts. This adds richness to the experiment, but also presents new difficulties in finding and aggregating a statistic that describes the simple consistency averaged over the entire group of artifacts.

In international comparison experiments, perceived complexity can increase dramatically as the amount of data increases. With conventional techniques, one must find a consensus model for a more intricate "right answer"; a consensus fitting protocol and perhaps even agreement about which measurements are to be regarded as "outliers" to be excluded from the fitting. As the experiment becomes more intricate, it becomes more challenging to create the "consensus right answer" as is traditionally required to begin a consistency analysis.

In contrast, agreement about essentially simple "consensus invariants" can be much easier to obtain. For example, two-dimensional coordinates of the balls of a ball plate, measured by any laboratory, could be used to calculate the distance between a specified pair of balls. All laboratories would be expected to report the same scalar value for this distance, within the dispersion of the reported combined uncertainty. In a typical ball plate, there would be an enormous number of distances to consider, and the random variables describing the distances would not all be independent of one another. All the resulting increase in arithmetical complexity can be dealt with by computers, to quantitatively describe the conceptual simplicity of the scalar "distance", averaged over all balls.

This paper discusses how the apparent complexity of comparisons can be addressed using a variant on traditional chi-squared testing. It employs the root-mean-square of the normalized differences between pairs of measurements, or consensus-invariant scalars. It avoids any necessity for a "consensus right-answer". It can create a conceptually simple description of average consistency, even in the richest of comparisons, at the expense of arithmetical complexity that is easily handled automatically by computers.

2. COMPARISON CONSISTENCY TESTING

The MRA process for reporting key comparisons is usually in terms of a key comparison reference value (KCRV), a good but not necessarily the best representation of the SI value. The KCRV is usually derived from the results of the participating NMIs considered as peers. When the KCRV is taken to be the weighted mean of the N values, classical chi-squared testing can be rigorous.

In cases where some other method (such as the simple arithmetic mean or the median) is chosen to represent the KCRV of the peer results, chi-squared testing must be extended to assess consistency [4,6].

Where non-Gaussian distributions are reported, including the Student distribution implied by the use of finite degrees of freedom, chi-squared testing must be extended further [4,6] to address the specific forms of the claimed distributions.

Chi-squared statistics have been extended to cover both of these situations, and can be used to test consistency against any choice of KCRV, even when the distributions associated with the reported uncertainties are not Gaussian. Such testing rigorously expresses the probability of a key comparison's chi-squared statistic being exceeded by the randomness in the resampled chi-squared expected from the claimed uncertainties. For some key comparisons there has been a significant delay in agreeing on a final report. Accepting extended chi-squared statistical testing should eliminate this delay when there is no compelling evidence that anomalies exist, so that neither additional causes of dispersion (outliers, biases or uncertainty components) nor other KCRV candidates need to be considered for the final report.

3. CLASSICAL CHI-SQUARED ANALYSIS

The classical chi-squared approach starts with the inverse variance weighted mean as the locator of central tendency used as the KCRV. Under the assumptions of a stable circulating artifact, and independent Gaussian uncertainty distributions, this weighted mean is suggested as the first choice for the KCRV [3]. For variances away from the weighted mean, the classical chi-squared statistic expressed as a reduced chi-squared with $(N - 1)$ degrees of freedom (note that its value is expected to tend towards 1 as N increases), is

$$\chi_c^2 = (N-1)^{-1} \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{u_i^2} \qquad (1).$$

If the N results are drawn from N independent Gaussian distributions, with the same unknown mean, and with standard deviations equal to the reported standard uncertainties, then the number $y_c = \chi_c^2$ computed for the particular key comparison data, $\{x_i, u_i\}$, is expected to be a sample of a positive random variable y distributed as an exact reduced chi-squared with $N{-}1$ degrees of freedom, with a probability density function (PDF) that is proportional to $y^{(N-3)/2} \exp(-y\,(N{-}1)/2)$.

Non-Gaussian distributions and inter-laboratory correlations can lead to departures from this exact chi-squared form [4,6], but the 'chi-squared-like' distributions for any given key comparison are readily evaluated by Monte Carlo simulation of the stated uncertainties. Chi-squared testing uses the fraction of this PDF that is greater than y_c, $P(y > y_c)$.

For any comparison, the classical chi-squared χ_c^2 has a particular value y_c and the probability $P(y > y_c)$ can be obtained from the chi-squared tables with the appropriate degrees of freedom, from packaged functions, from numerically integrating the analytic form of the PDF given above, or from a Monte Carlo simulation for this particular comparison. P($y > y_c$) gives the probability that y_c would be exceeded by chance, again given that all N results are drawn from N independent Gaussian distributions, with the same unknown mean and with standard deviations equal to the claimed standard uncertainties. This is the classical null hypothesis test for a group of peer results and is proposed [3] as a necessary test for not rejecting the weighted mean as the KCRV. The cumbersome language is an unfortunate byproduct of the fact that experimentally, evidence can only be rigorously compelling for rejection. The proposed threshold [3] is that if the theoretical chance of exceeding y_c is less than 5%, then the comparison should be regarded as inconsistent with agreement to the weighted-mean KCRV within the claimed uncertainties.

This means that one KC in 20 would be expected to be unjustly rejected. Even this high level of rejection does not in any way create a high level of confidence that the KCRV model is correct, and actions based on acceptance of the KCRV model need to be viewed in this light.

With appropriate Monte Carlo simulation, the same type of test can also be used to examine other values of central tendency that might be used to represent the KCRV. For example, as a KCRV candidate, it is possible to consider the median [4,6], for which null hypothesis testing using chi-squared statistics is traditionally regarded as too difficult.

Lab Name	*Lab Submitted Value* /nm	*Lab Submitted* $u(k=1)$ /nm	v_S
NRC	29	14	10
CENAM	45	7	82
INMETRO2	81	14	13
INMETRO1	42	2	5
INTI	40	11	78
NIST	59	10	10000
CEM	47	8	74

Table 1: *Key Comparison data set from CCL.SIM-K1 short gauge block calibration by optical interferometry. The data set includes the submitted measured value, standard uncertainty and degrees of freedom for the steel gauge block of nominal 8 mm length.*

	Conventional χ^2	*Extended* χ^2
Weighted Mean	43.1 nm	43.1 nm
Uncertainty	1.5 nm	1.5 nm
Experimental $\chi^2 = y_c$	1.975	1.975
P($y>y_c$)	6.5 %	10.2 %

Table 2: *Values for weighted mean for Table 1, its associated uncertainty and* P($y>y_c$) *evaluated using conventional statistics (Gaussian distributions), and using Monte Carlo simulation to evaluate the probability for extended chi-squared statistics that incorporate the claimed degrees of freedom* v_S*.*

Table 2 lists the values for the weighted mean (specifically, the weighted mean with weights proportional to the inverse of the square of the standard uncertainty), its associated uncertainty and the probability of the chi-squared statistic exceeding the experimental value y_c by chance. It is calculated for the Gaussian approximation (see Figure 1) which corresponds to the conventional chi-squared test. The extended chi-squared-like probability, that takes account of the claimed degrees of freedom, was calculated by Monte Carlo simulation of the comparison-specific distributions (see Figure 2).

The differences between Figures 1 and 2 initially appear to be rather subtle. They are the result of taking into consideration the tails of the Student distribution; a reflection of the larger uncertainty in the uncertainties stated by participants with low degrees of freedom. Figure 2 takes into consideration the degrees of freedom submitted by each participant. Despite the apparent subtlety, the probability of exceeding the experimental chi-squared is substantially increased for the extended chi-squared testing as shown in Table 2. This simple example illustrates the necessity and advantages of including the degrees of freedom, as claimed by the participants, when evaluating their consistency. The degrees of freedom may reflect a lab's uncertainty in their uncertainty evaluation, and even here there is a rigorous basis for using the Student distributions in this way. Monte Carlo resampling offers a rigorous statistical method to incorporate all GUM-compliant participant data in consistency and equivalence testing.

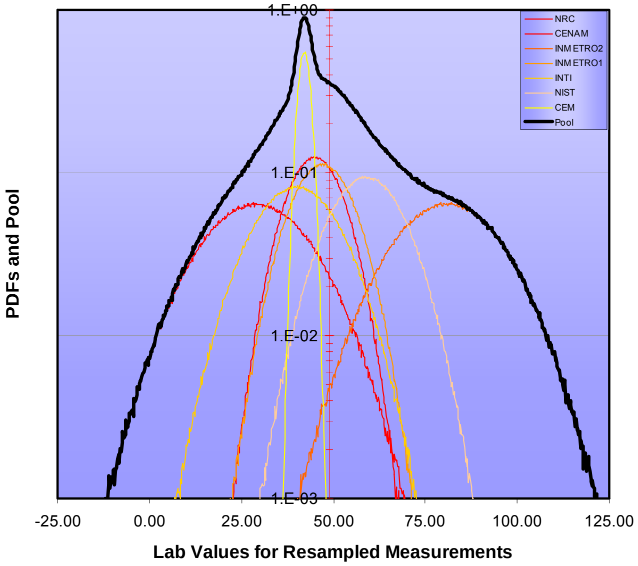

Fig. 1: *Resampled Lab Values' PDFs and pool for the 8 mm steel gauge block plotted as normal distributions.*

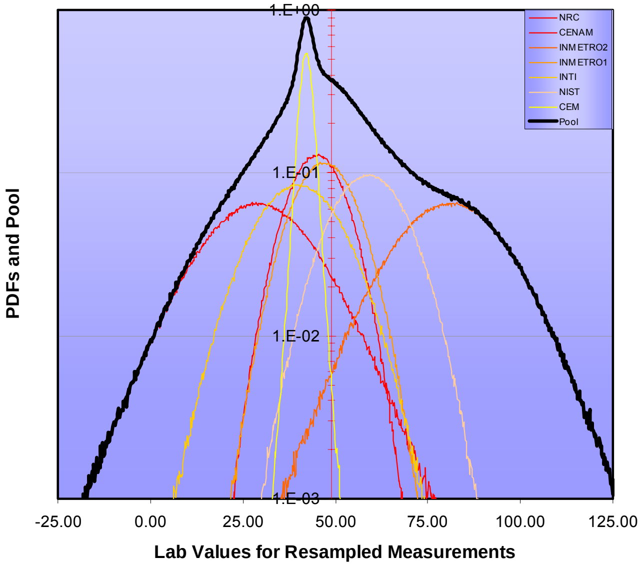

Fig. 2: *Resampled Lab Values' PDFs and pool for the 8 mm steel gauge block plotted as Student distributions, taking into consideration the reported degrees of freedom.*

The subtleties in the wings of claimed uncertainty distributions can change the interpretation, for a substantial range of possible experimental y_c's, (y_c between 2.1 and 2.4 in this case) and avoid the problematic conclusion of "null hypothesis rejected". Doing the analysis properly, to account for the subtleties, can "rescue" a comparison from unnecessary doubts and delays.

4. χ^2 AVERAGING TO REMOVE COMPLEXITY

Conceptually, consensus invariants of a comparison can be much simpler to describe ("the labs are expected to obtain the same results for…") than a complete fitting model for an artifact that embeds this expectation of invariance. To test whether the labs *did* get the same results, within the range to be expected from their claimed uncertainties, can be very simple. To do this we can compare a particular scalar consensus invariant, obtained by pairs of labs, normalizing the difference to the expected combined standard uncertainty in the difference, and doing a mean-square average

$$\chi^2_{APD} = [N(N-1)]^{-1} \sum_{j=1,\neq i}^{N} \sum_{i=1}^{N} \left[\frac{(x_i - x_j)}{u(x_i - x_j)} \right]^2 \qquad (2).$$

This all-pairs-difference chi-squared [2,5,6] can be further averaged, for example over *all* the consensus invariants of a comparison. Note that it is also the mean-square aggregate of the simplest normalized error or E_n [7].

The chi-squared-like statistic of Equation 2 can be evaluated by Monte Carlo simulation [5] in much the same way as was illustrated above for CCL.SIM-K1. The essence is to simulate the χ^2_{APD} by Monte Carlo simulation of the claimed uncertainties, but constrained by the null hypothesis of perfect agreement for each consensus invariant. In this way, not only the degrees of freedom as claimed by each participant, but also the overall "degrees of freedom" for the aggregated chi-squared-like statistic can be fully accounted for – including all consequences of the fact that the consensus invariants are not generally statistically independent. No "counting" of degrees of freedom is necessary, since all of the relevant details are handled by the Monte Carlo simulation.

Building a Monte Carlo model of the claimed uncertainty budgets, as applied to the consensus invariants, can be very easy but can also be challenging when there are approximations in the claimed uncertainty budgets that need to be undone to obtain consistency over the whole range of the consensus invariants.

Table 3 shows the KCRV-free analysis of consistency using the statistic of Equation 2. The deficiencies of the KCRV-mediated conventional χ^2 test are evident: with Gaussian uncertainty distributions this comparison should be deemed to have sufficiently compelling evidence to ***reject*** the null hypothesis of agreement of paired results. Where a pair of results are somewhat far from the KCRV, the KCRV-mediated conventional χ^2 cannot reliably distinguish when the pair agrees well with each other and when the pair of results are equidistant from the KCRV but lie on opposite sides of the KCRV [5].

When the degrees of freedom claimed by the participants is properly accounted for, this comparison is “rescued” when the 5% rule [3] is applied to the statistic of Equation 2.

	Gaussian	*Extended, Student*
Experimental $\chi^2_{\mathrm{APD}} = y_c$	2.181	2.181
P(y>y_c)	4.9 %	8.1 %

Table 3: *Values for χ^2_{APD} for Table 1, and* P(y>y_c) *evaluated by Monte Carlo simulation for Gaussian distributions, and Student uncertainty distributions that incorporate the claimed degrees of freedom ν_S.*

This analysis is independent of any choice of KCRV or fitting model. It addresses the essential metrology of a comparison: whether results, that are supposed to be the same, *are* the same within the variation expected from the claimed uncertainties. The metrological prediction that they should be the same can be tested, by the classical scientific method of prediction and experimental validation, in very broad aggregates to elevate the metrology in question to the level of a true measurement science.

5. CONCLUSIONS

Simple answers to the question of agreement between NMIs are best addressed by exploiting all of the richness of the measurements performed, and data collected, during the KC experiment. This is the primary motivation for employing unmediated chi-squared-like testing for the evaluation of statistical consistency of a KC data set. The evaluation and defense of uncertainty budgets remains an essential priority of the metrology community since conclusions regarding intra-laboratory consistency depend on the description of the probability density function associated with the uncertainty claims. For this reason, it is important to weigh the consequences of relying on low-probability events (which produce the ‘tails’ of the distribution associated with the uncertainty budgets) for making decisions on who agrees with whom.

REFERENCES

[1] www.bipm.org

[2] Steele, A.G., Hill, K.D., Douglas, R.J. "Data pooling and key comparison reference values", *Metrologia* **39** (3), 269-277, June 2002. doi:10.1088/0026-1394/39/3/4

[3] Cox M. G., “The evaluation of key comparison data” *Metrologia*, **39** (6), 589-595, December 2002. doi:10.1088/0026-1394/39/6/10

[4] Steele, A.G., Douglas, R.J. "Chi-squared statistics for KCRV candidates", *Metrologia* **42** (4), 253-261, August 2005. doi:10.1088/0026-1394/42/4/009

[5] Douglas, R.J., Steele, A.G. "Pair-difference chi-squared statistics for Key Comparisons", *Metrologia* **43** (1), 89-97, February 2006. doi:10.1088/0026-1394/43/1/013

[6] Steele, A.G., Douglas, R.J. "Extending chi-squared statistics for key comparisons in metrology", *Journal of computational and applied mathematics* **192** (1), 51-58, July 15, 2006. doi:10.1016/j.cam.2005.04.041

[7] Steele, A.G., Douglas, R.J. "Extending E_n for measurement science", *Metrologia* **43** (4), S235-S243, August 2006. doi:10.1088/0026-1394/43/4/S10