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

CONSISTENCY MEASURES FOR PEER-TO-PEER COMPARISONS

A. G. Steele [1], *R. J. Douglas* [2]

Institute for National Measurement Standards, National Research Council, Ottawa, Canada
[1] alan.steele@nrc-cnrc.gc.ca
[2] rob.douglas@nrc-cnrc.gc.ca

Abstract: This talk summarizes statistical approaches to consistency analysis using key comparison data, including conventional mediated testing relative to the inverse variance weighted mean, extensions of this chi-squared methodology to accommodate finite degrees of freedom in the uncertainty claims, and a fully bilateral approach that avoids invocation of any key comparison reference value.

Keywords: key comparison, chi-squared test, consistency.

1. INTRODUCTION

International comparisons have stable artifacts circulated and measured by different National Metrology Institutes (NMIs). These comparisons are usually conducted in a "peer to peer" style, where the "true value" is treated as an unknown, and a "reference value" may be chosen using the measurements from the participating NMIs: treating them, without prejudice, as peers.

Key Comparisons are organized under the auspices of the CIPM and published in the BIPM Key Comparison Data Base (KCDB) in support of the Mutual Recognition Arrangement. Evaluation and reporting of consistency in Key Comparisons has usually been done after the selection of a Key Comparison Reference Value (KCRV). The KCRV is usually chosen by a clearly described method using the peer results reported by the participants.

After a KCRV has been chosen, consistency relative to that KCRV can be described in a variety of ways. The degrees of equivalence of a participant relative to that KCRV are given in the KCDB as the departure of the participant's value from the KCRV, and the 95% coverage interval for that departure. The ratio of these is closely related to the traditional individual measure of consistency, the normalized error E_n – and the 95% coverage factor k. The participant's degrees of equivalence ratio (departure : expanded uncertainty) is E_n/k, and is closely related to other measures of consistency.

The reduced χ^2 is another powerful measure of the overall consistency for a comparison where the participants have reported independent Gaussian uncertainties and the KCRV method is the inverse-variance weighted mean of the participants' values. We have extended these methods to encompass results reported with effective finite degrees of freedom and to other methods for choosing the KCRV (such as the median) [1]. The extended reduced χ^2 statistic is the mean-square E_n of the participating laboratories, which is distributed as a chi-squared-like distribution that can be evaluated for any specific comparison by Monte Carlo simulation of the participants' claims that have been constrained by the null hypothesis of ideal agreement of all the participants' means with the KCRV. Although the reduced χ^2 can be calculated for a single participant, in this role it offers no striking advantage over E_n as a measure of consistency: it merely suppresses the sign. However, in many cases it has a familiar shape of distribution, and can approach a reduced chi-squared with one degree of freedom.

2. DISCUSSION

Traditionally, chi-squared statistics are used to exploit their ability to aggregate a measure of consistency across the N participants in a comparison. Under the assumption of ideal agreement of the participants' underlying (but unknown) means, the resulting chi-squared statistic again has a familiar distribution in many cases: a reduced chi-squared with (N-1) degrees of freedom. Larger numbers of participants are associated with sharper distributions that imply less arbitrariness in the reduced χ^2 value, and for decision-making based on reduced χ^2 analysis, the appropriate distribution is used to account for the arbitrariness with a carefully constructed statement of probability (for example "…with the participants' claimed uncertainties, assuming ideal agreement of the participants' mean value gives a χ^2 statistic that exceeds the experimental χ^2 value of this comparison with a probability of x%...").

These measures of consistency are not metrics, but are closely related to a statistic that *is* a metric (also known as a norm): the Root Mean Square E_n. For N participants, the RMS E_n is a metric for the consistency of their values' agreement with this KCRV, to within the claimed uncertainties. All of the usual intuitive ideas about ordering apply to agreement measured in this way, but may not apply to measures that are not metrics. In addition to calculating this metric – in much the same way as can be done for a χ^2 statistic – it is possible to evaluate the probability of the

claimed uncertainties generating an RMS E_n, larger than the experimental RMS E_n, under the constraint of the null hypothesis that all laboratories have a common mean that can be represented by the specified method's KCRV. These provide a common basis for rigorous hypothesis testing that should appeal both to those familiar with chi-squared testing, and to those familiar with E_n tests.

These measures of consistency have also been recommended for application in a Key Comparison during the quest for an appropriate method for determining its KCRV, for testing KCRV methods and even for choosing a consistent subset of participants. Different methods can be evaluated as candidates for choosing the KCRV, and even a defined procedure for selecting a suitable method can have its full decision tree evaluated to determine the probability of mistakenly labeling a comparison as "not consistent". In the search for a consistent subset, "outliers" are identified and excluded from the determination of the KCRV. We note that for the N dimensional metric space of the RMS E_n, outlier rejection is not simply described by projection into a subspace but requires both a projection into a subspace (where the outlier's departure from the KCRV is forced to zero) and the renormalization of the RMS E_n by $(N/(N-1))$. Outlier rejection also changes the KCRV. Processes for searching out a consistent subset (by choosing different outliers) will be discussed in the context of jumping from subspace to subspace, and the intricacies of evaluating these decision trees will be presented.

Rather than use measures of consistency that are mediated by the KCRV, we advocate using unmediated measures of consistency that exactly emulate the simplest use of measurements from different participants, based on the RMS of all the relevant bilateral E_n. Unmediated pair consistency can be aggregated in a large number of meaningful ways: one participant with respect to all other for a single artifact; or for all participant pair for a single artifact. Each of these styles can be aggregated over multiple artifacts in the same comparison; or aggregated across a scale; or aggregated across different methods; or aggregated across different measurands. For all these, pair consistency is an easily explained concept that matches the simplest use of measurements – comparing one measurement directly to another – without mediation (needing neither the "right answer" nor a reference value nor a designer's tolerance interval).

For example, the KCDB can be "mined" to determine an RMS E_n for a particular NMI that measures its average consistency with another specific NMI – or with "any other" participating NMI, by root-mean-square averaging of the pair E_n in each comparison over all other participants. These broad averages are proposed as useful measures of consistency when NMIs are called upon to demonstrate the efficacy of their quality systems. The overall RMS average over all Key Comparisons would be a quantitative measure of the consistency of the KCDB's participants relative to the SI paradigm. The advantages of unmediated consistency measures will be presented, and discussed in the context of outlier rejection and searches for the largest consistent subsets.

Pair consistency provides a powerful means of searching for the "largest consistent subset", again *before* a KCRV is chosen. It is a parameter-free discrete search that can be done very efficiently by ranking the RMS E_n and considering the largest value to be the best candidate for "outlier" status. If multiple outliers are to be identified, the issue of whether or not this is the "globally most consistent" set of participants can easily be addressed. For comparisons with fewer than N=20 or 30 participants, it is feasible to evaluate all 2^N combinations and calculate the pair consistency from the RMS pair E_n for all $[2^N-(N+1)]$ possible subsets with 2 or more participants.

3. SUMMARY

The subtle distinctions between consistency-with-a-particular-KCRV-method and pair-consistency will be discussed. Demonstrating pair consistency will be discussed as a necessary condition for any demonstration of consistency in utilitarian metrology. We will discuss the deficiencies of relying wholly on testing consistency-with-a-particular-KCRV-method. Pair-consistency tests will be presented for the screening of comparisons, before a KCRV is chosen. Strategies for evaluating probabilities of mistakenly rejecting the null-hypothesis will be outlined. Through all this we will emphasize the distinction between non-rejection of the null hypothesis (common in metrology) and the ultimate acceptance of the null hypothesis (which is almost never done in international comparisons).

Methods will be discussed for the efficient extension of chi-squared-like consistency tests to situations where there are thousands of pooled comparisons under consideration (such as the KCDB). Ab-initio regeneration of all the comparisons is not necessary when a new comparison is to be added to the pool – if the proper information is preserved, the simulation of the new comparison can be incorporated into the previous pool and the required chi-squared-like probability density function can be obtained by the properly scaled numerical convolution of two histograms derived from Monte Carlo simulations. The choice of appropriate scaling is discussed.

REFERENCES

[1] A. G. Steele R. J. Douglas, "Chi-squared statistics for KCRV candidates", *Metrologia*, **42**(4), 253-261, 2005.

[2] A. G. Steele, K. D, Hill, R. J. Douglas, "Data pooling and key comparison reference values," *Metrologia*, **39** 269-277, 2002.

[3] A. G. Steele, R. J. Douglas, "Extending chi-squared statistics for key comparisons in metrology", *Journal of computational and applied mathematics*, in press.