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

ANALYSIS OF MEASUREMENT COMPARISONS

Alan Steele

NRC, Canada

Q1: "Can we use the probabilistic interpretation of uncertainty budgets when analyzing measurement comparisons for consistency among the participants?"

Measurement comparisons are the simplest complete element of measurement science, and the language of probability is a candidate for standardizing communication about the consistency of measurement comparisons. A measurement's true worth is only harnessed by comparison to other measurements, or to summaries of previous measurements (such as tolerances). Carefully designed and executed comparisons are analyzed to provide confidence about consistency within the claims of the uncertainty budgets. This confidence may be quantified with probabilities.

Focusing on comparisons as the ultimate goal of metrology creates opportunities for clarity. Descriptions of knowledge about a measurand have distractingly different approaches. Some are widely discussed, and others are less so: there are different schools of statistics, and each provides its own approach to describing measurement uncertainty – usually with some approach to probability. The frequentist and Bayesian schools are widely known, but there is also the fiducial approach, the marginal likelihood approach, the probability/possibility approach of the Dempster-Schaffer theory, and the predictive approach used in experimental sciences and spectacularly vindicated in the development of quantum mechanics and stringent tests such as Bell's inequality. The uncertainty budget, whether formulated as a description in accordance with statisticians' views, or as a prediction in terms of physicists' views, *must* be useable as a prediction to use the scientific method and recover the clarity provided by a comparison.

Fortunately, the ISO GUM provides a standard framework that can encompass many approaches to expressing uncertainty in measurement. The ISO GUM also provides the basis for *using* measurements and their uncertainties, for example in a comparison, by creating and treating a new measurand: the difference between two participants, or between a participant and a reference value. The power of the ISO GUM approach rests on its treatment of uncertainties with distributions of probability, and in the context of a comparison these distributions can be treated as predictions that are subject to experimental testing – this is vital since it is the only test that can justify metrology's aspiration to be a rigorous measurement science.

The simplest example in the ISO GUM is the probabilistic calculation of a coverage factor and the associated expanded uncertainty. Note that, by the procedures given in the ISO GUM, *expanded uncertainties are not used* in any subsequent calculations – if reported, they are to be converted back to standard uncertainties for use in subsequent calculations. Using the probability distribution of the combined uncertainty, the generalization of the coverage factor is an integral of the probability multiplied by an appropriate cost function.

The calculus for uncertainty distributions can be done by approximations such as the "law of propagation of uncertainties" for Gaussian distributions, or – is with the aid of the Welch-Satterthwaite approximation – for Student distributions (associated with claims of finite degrees of freedom), by analytic or numerical integration over distributions carried through the measurement equation, or by Monte Carlo simulation to combine the simulated randomness of the input quantities through the measurement equation to the randomness predicted for the output quantity.

1. What are the uses of explicit probability distributions for uncertainties?

2. Should the sense of "probability" be more carefully defined as a "claim" and/or as a "state of knowledge"?

3. What role should be assigned probability statements associated with statistical aggregates of metrological consistency?

Q2: *"Can statistical data analysis help simplify our understanding of comparison results, including the notions of a reference value, agreement, and for MRA KC only, 'degrees of equivalence' among the participants?"*

For multi-participant comparisons, the pilot laboratory and the other participants are naturally alert to the possibility of the presence of additional sources of uncertainty, beyond those accounted for in their uncertainty budgets. When this is so, they strive to understand and to control this new effect. However, this can lead to unconscionable delays in the publication of results, which might be avoided with a simpler report on the observed dispersion of results. Should statistical methods be used to help attribute a source for unexpectedly large differences of results – treating one result (or more than one result) as an outlier, or attributing misbehaviour to one (or more) of the circulated artefacts?

When no compelling evidence is found for any additional sources of uncertainty, one might hope for speedy publication of the comparison. However, even in these cases it is common for the final report to take a year or more to prepare and often the delay has been associated with reaching agreement on a reference value. Can alternate statistical means be used that more simply reflect the actual usage of the constituent measurements? Can alternate statistical methods help to speed initial publication in these cases?

Even when a comparison exhibits surprisingly good agreement, with the observed dispersion being much less than was predicted by the participants' uncertainty budgets, the time taken for the initial publication is not always short. Can statistical tools help identify these cases as candidates that require fast-track publication?