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

# HARNESSING SIMULATIONS FOR GUM-COMPLIANT EVALUATION OF UNCERTAINTY IN MEASUREMENT

J. E. Decker [1], A. G. Steele [2], R. J. Douglas [3]

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

**Abstract:** This paper presents a specific implementation suite and some background for combining uncertainty components, and using them, through the propagation of distributions using Monte Carlo simulation.

**Keywords:** Monte Carlo, probabilistic interpretation.

## 1. UNCERTAINTY BUDGETS AND MONTE CARLO

The ISO *Guide to the Expression of Uncertainty in Measurement* [1] proposes Monte Carlo simulation as a technique for combining uncertainties when simpler methods may not apply. Monte Carlo simulation as the reference technique is presented in some detail in its Supplement 1, which has been circulated widely, starting with its draft form in 2004 [2].

In this paper we demonstrate the simplicity that may be obtained by adopting Monte Carlo techniques for GUM-compliant combination of standard uncertainties. We discuss what additional information needs to be presented so that anyone – from neophyte to professional statistician – can have confidence in the results of a Monte Carlo simulation as a proper representation of the consequences of the claimed uncertainty components.

## 2. END USES FOR UNCERTAINTY BUDGETS

Universally, the underlying purpose for statements of uncertainty is to guide decisions involving the measurand. The metrologist's uncertainties are propagated through one or more traceability chains to arrive at the decision, which conventionally is based on a designer's tolerance interval. When we recall that the designer's experience is *also* based on measurements, it is easy to appreciate that the weakest part of the overall chain (relating a metrologist's result to a designer's measurement-based experience) may well be the conventional tolerance interval. More sophisticated designers might express past experience as a "cost function" rather than as a uniform tolerance interval, and want access to all details of the metrologists' uncertainty distribution. Monte Carlo simulation, applied to uncertainty analysis, can support these more sophisticated approaches that express more detail about a designer's requirements.

For coupling to a designer's cost function that embodies commercial experience (likely proprietary), the uncertainty may be best captured by an uncertainty budget as expressed in a Monte Carlo model that accompanies the measurement. Other ways of providing this information are also possible: as moments of the as-simulated distribution or as a histogram representing the as-simulated distribution. The uncertainty budget, expressed as a Monte Carlo model, has a clarity that greatly simplifies the proper identification and treatment of effects that are fully covariant between the measurement and the measurements that contributed to the designer's experience.

Usually, the role for Monte Carlo simulation is in support of treating uncertainties using techniques such as the "law of propagation of uncertainty" (LOPU) [1, Clause 5.1.2]. This support is important when the validity is not self-evident for linearized approximation of the measurement equation, or the Gaussian approximation of uncertainty distributions associated with particular causes, or of the Welch-Satterthwaite approximation.

When there is a consensus about the general target accuracy for the treatment of uncertainties, Monte Carlo simulation can be the easiest approach to validating these approximations for the specific case under consideration. For example, in high-level metrology, the authors are unaware of any application requiring uncertainties in measurement to be determined to better than 10%, and at these levels, the demonstration by Monte Carlo simulation of the adequacy of the GUM's approximations can be very straightforward.

In any case, the end use of the measurement and its uncertainty can provide invaluable guidance about the appropriate use of advanced techniques such as Monte Carlo simulation.

## 3. UNCERTAINTY CLAIMS AS PREDICTIONS

For measurement science to sustain the confidence of the physical sciences and engineering, we must be seen to be using the scientific method: comparing prediction with experiment. Although there are many ways of doing this, the closest fit to any full chain of metrological inference is found by comparing pairs of measurements that have been made on nominally identical measurands.

In this essential test, the uncertainty budgets for the pair are treated as the definitive claim about nominal identity and on which the probabilistic prediction is based, with the measured results as the definitive experimental results. The method for combining uncertainty components is not constrained (LOPU, LOPU with the Welch-Satterthwaite approximation, Monte Carlo simulation, probability calculus…) nor is the school of statistics (frequentist, marginal likelihood, Bayesian, fiducial …) used to help one arrive at the standard uncertainty for the pair difference. The chosen method's standard uncertainty, of the pair difference, serves as the definitive parameterized prediction.

We see advantages in giving a specific name to the ratio: the *difference* between one measured result $x$ and another result $x_r$ (often $x_r$ would be a reference result), $(x-x_r)$, divided by the *standard uncertainty of that difference*, $u(x-x_r)$.

$$(x-x_r) / u(x-x_r)$$

At first glance, this ratio may appear deceptively familiar, particularly to non-metrologists. We have proposed [3] naming this quantity "the metrologist's ratio" to help distinguish it from its close relatives.

In defining the uncertainty scale factor, the GUM-compliant usage of the "standard uncertainty" places some specific constraints on the more general concepts of normalized error, $E_n$; or of the Z-score. For GUM compliance, neither of these are sufficiently specific terms: $E_n$ is often applied with an expanded uncertainty in place of the standard uncertainty, and there are other meanings to the term Z-score.

Furthermore, the ISO GUM specifies the standard uncertainty to be one of two different things: either the standard deviation of the predicted distribution, OR (if a degrees of freedom parameter is specified) as the experimental standard deviation associated with the scaled and shifted Student-*t* distribution with the specified degrees of freedom, $\nu$. For $\nu>2$, the square of the GUM's standard uncertainty is smaller than the variance of the Student-*t* distribution by the factor $(\nu-2)/\nu$. In the context of pair differences, the combined standard uncertainty is specified even more tightly by the ISO Guide – as the combined uncertainty with its effective degrees of freedom calculated using the Welch-Satterthwaite approximation for the linearized measurement equation. The "metrologist's ratio" comfortably encompasses all these subtleties in a way that a more general term cannot.

The metrologist's ratio is a straightforward statistic for comparing measured with claimed agreement. With a probabilistic analysis of the ratio (or aggregates of multiple instances of this kind of ratio) using the uncertainty claims, we can test the null hypothesis (of agreement of the means) and the measurement equation (of the difference, usually assembled as the difference of the two measurement equations). Monte Carlo simulation is an excellent tool for doing this, since for this test we want to use all information available in the uncertainty budgets to understand the significance of the measured differences relative to the full uncertainty claims.

Demonstrating the scientific method in this way should circumvent all of the controversies that can arise about the meaning of what is being simulated: here the claimed randomness is being tested, along with the null hypothesis (of agreement), relative to the experimental measurements that are available. The bitter experience of those charged with teaching statistics is that simulations can be misapplied, and initial skepticism is to be expected particularly from any statisticians who are unaccustomed to seeing statistics as part of what is being judged by experiment. We see the statistical prediction of a standard uncertainty as no less subject to experimental validation than would be the probabilistic predictions from quantum mechanics. Notice that we are always simulating the claimed uncertainties, and never need to try to simulate the unavailable "true" distributions associated with the uncertainties.

Some, but not all, established pseudo-random number generators have all the desirable properties for uncertainty evaluation. In order to quantify improbable eventualities, there are fairly stringent requirements on how the parent uniform generator creates pseudo-random numbers near 0 and near 1, since these are used for creating the tails to the left and to the right of the resampled pseudo-measurements. Not all parent uniform generators treat these alike, and this needs to be tested and fixed if there is an asymmetry evident at the dynamic range that will be used in the evaluation of uncertainties. To exploit Monte Carlo simulation for uncertainty evaluation, we will need to develop and communicate a confidence in the pseudo-random number generators and models that are used.

## 4. GRAPHING FOR CONFIDENCE

We illustrate the power of data display in a histogram format to help achieve this confidence. Recall that the model equation of the measurement assigns a value to the output quantity based on a set of values of the input quantities. For each uncertainty component of an input quantity, the histogram of the simulation's re-sampling can be compared with the analytic curve claimed for the probability density function of this component. The output quantity's histogram, as determined for each resampling by the model equation, can also be usefully compared with analytic approximate forms.

In Figure 1 we show the graph of an analytic Gaussian or normal distribution (recall that the logarithmic graph of a Gaussian is a parabola opening downwards), along with histograms derived from the Box-Muller algorithm [6] operating on $2\times10^{10}$ events from two different parent uniform generators: one (in grey) is for the Hill-Wichmann generator [7] with a dynamic range of about 15 bits, and the other is for the RANLUX64 generator, with excellent theoretical properties [8] with a dynamic range of 48 bits. Figure 1 shows that, for the context of the Box-Muller method, the dynamic ranges for simulating a Gaussian are easily appreciated as being almost $10^5$ and at least about $10^8$.

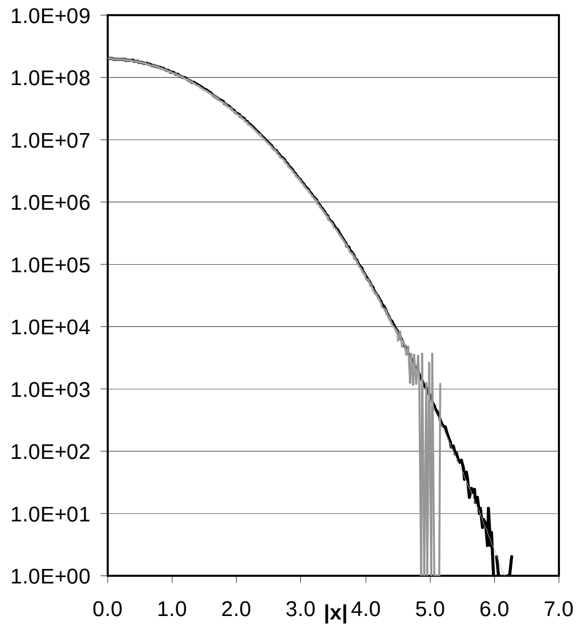


Figure 1. Histograms comparing $2x10^{10}$ simulated events of a standard normal or Gaussian, implemented by the Box-Muller algorithm from two different parent uniform generators: the Hill-Wichmann generator (grey) and the RANLUX64 generator (black). The analytic Gaussian curve is indicated by the pale grey hairline.

Graphs such as those in Figure 1 can convincingly communicate how well the pseudo-random generator can recapture the analytic claim that is typically made in an uncertainty budget. The correlation characteristics of current pseudo-random generators are usually more than adequate to represent the claimed independence of resampled uncertainty components from the different uncertainty budgets under consideration.

## 5. MEASUREMENT VERSUS MEASURAND

Simulations can easily deal with skewed distributions as well as the more usual symmetric ones. If asymmetric uncertainty distributions are being used, additional precautions are necessary. Care is needed to assign the correct sign of skewness to the uncertainty distribution of the measurand, which is reversed in sign from the skewness of a distribution of measurements [4]. These precautions are not fundamentally different than those needed for applying a static correction with the correct sign: if some physical cause *increases* a measured value, then the measurand value (corrected for this effect) must be *below* the uncorrected measured value and the absolute value of the correction must be subtracted. There are simple rules for simulations involving skewed distributions, with even simpler rules for dealing with symmetric distributions:

1) Collect the information about the input quantities' distributions, noting carefully whether they are asymmetric.
2) For asymmetric distributions, carefully determine whether the distribution is GUM-compliant, describing where the measurand is expected to be (with respect to the reported value) or if the distribution is describing where repeat measurements are expected to be (with respect to the reported value).
3) Do not alter asymmetric measurement distributions, but resample from these directly.
4) By reflection [4] in the reported value, convert asymmetric measurand distributions into the measurement distributions of the correct skewness, and resample these by Monte Carlo simulation, feeding the results of each simulated input quantity into the measurement equation to obtain the resampled output quantity (note that the measurement equation itself may have skewness-altering properties).
5) Repeat the resampling to create a histogram of the output quantity as a measurement distribution.
6) Decide on the value that would be reported to represent this distribution (this may well be the mean, but it is to be the value that you would report)
7) Reflect this histogram of the resampled output quantity in this value to obtain the GUM-compliant distribution, describing where the output quantity measurand is expected to be (with respect to the reported value).

This straightforward change of variable can be used to transform between the two types of probability distribution functions (PDFs) used in metrology: the PDF of the measurand and the PDF of the measurement results. In assembling asymmetric PDFs, these transformations can help keep the symmetry of each distribution properly aligned, even for asymmetry that is merely caused by uncorrected biases. This simulation may be done to quantify the effects of asymmetry on the final PDF and its coverage intervals, which may often be used as a justification for ignoring the asymmetry.

## 6. MONTE CARLO IN EXCEL

We provide a ‘toolkit’ for performing the basic Monte Carlo [9, 10] techniques proposed in *Supplement One* [2], using Visual Basic for Applications, the built-in programming language that is supplied with the Microsoft Excel spreadsheet program. This Excel toolkit is useable and modifiable on an extremely large installed base of computers. It is fully open source and is supported by extensive comments, presentations and papers; with user instruction and training materials available on its website [9]. This means that the Monte Carlo resampling of an uncertainty budget can be passed on and used in subsequent steps of the chains on metrological inference. This can be helpful to end users who wish to do their own cost function calculations, or who want to use their own measurement equations and quantify the adequacy of simpler methods as applied to their own purposes.

The toolkit includes easy-to-understand functions for generating pseudorandom numbers distributed uniformly, or transformed as Gaussian or Student deviates as are commonly required in uncertainty evaluation. Emphasis is placed on comprehensibility over speed, since code validation and confidence in the propagated distribution method are paramount at this early stage of adoption. Although it should be rarely needed for speed of execution, the use of C language programs for macros in this Excel toolkit is also illustrated in some detail. Worked examples are used in the supporting documentation [9] to illustrate the impact of some of the approximations implicit in the GUM approach.

## 7. NULL HYPOTHESIS TESTING

Monte Carlo simulation is moving on to new roles as it is being broadly accepted as a reference method for propagating uncertainty claims. Probabilistic tests of the null hypothesis, based on the use of chi-squared-like statistics, aggregating the square of the "metrologist's ratio" discussed above, are enabled by simulation in the context of extended tests. These fundamental tests for metrological agreement within claimed uncertainties can be performed either as mediated tests using either an externally-supplied reference value, or a peer-determined reference value [11], or unmediated tests [12] using unmediated bilateral comparisons, with maximum re-use of the peer data, to exactly emulate the simplest practical use for the comparison data.

These new methods can test the consistency of the claimed uncertainties using a group of experimental values of one or more putatively invariant measurands. The Monte Carlo simulation can evaluate the probabilistic consequences of the randomness associated with the claimed uncertainties, and for a wide variety of circumstances it can provide a better estimate of the probability (of the claims exceeding the value of the experiment's statistic) than would be available from tabulated value of the chi-squared statistic. For many interesting cases, there are significant departures from the analytic chi-squared function in the sense that the use of an analytic chi-squared function would often misidentify a group of measurements as being inconsistent.

These new methods are aimed at transmitting confidence in the consistency of groups of measurements, and the familiarity of the chi-squared, or RMS $E_n$ (a norm [5]) will provide a prepared audience for statistical measures employing them. Their formal statistical efficiency remains to be studied by those interested in these aspects of statistics, but discussions of the formal statistical efficiency need to be tempered with a discussion of the metrological efficiency for broadly communicating confidence in consistency. For "metrological efficiency", we believe that familiarity is more important than the formal "statistical efficiency". Monte Carlo simulation can, of course, support unfamiliar statistics as well as chi-squared-like statistics, but at present we think it best to use the most familiar statistics.

## REFERENCES


[1] ISO *Guide to the Expression of Uncertainty in Measurement* 1995.

[2] BIPM Joint Committee for Guides in Metrology *Draft Supplement 1 to the Guide to the Expression of Uncertainty in Measurement* 2004.

[3] A.G. Steele and R.J. Douglas, "Simplicity with Advanced Mathematical Tools for Metrology and Testing" *Measurement* (in press) 2006.

[4] R.J. Douglas. A.G. Steele, B.M. Wood and K.D. Hill, "A Useful Reflection", *Metrologia* **42**, L35-9, 2005.

[5] A.G. Steele and R.J. Douglas, "Extending $E_n$ for Measurement Science", Metrologia **43** (in press), 2006.

[6] G.E.P Box and M. Muller "A Note on the Generation of Random Normal Deviates", *Ann. Math. Statist.*, **29**, 610-611, 1958.

[7] I.D. Hill and B.A. Wichmann, "An Efficient and Portable Pseudo-random Number Generator", *Appl. Statist.* **31**, 188-190, 1982.

[8] M. Lüscher, "A portable high-quality random number generator for lattice field theory simulations", *Computer Physics Communications* **79**, 100–110, 1994.

[9] http://www.inms-ienm.nrc-cnrc.gc.ca/qde/montecarlo/

[10] A.G. Steele and R.J. Douglas, "Simplifications from Simulations: Monte Carlo Methods for Uncertainties", *NCSL International Measure*, in press, 2006.

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

[12] R.J. Douglas and A.G. Steele, "Pair-difference chi-squared statistics for Key Comparisons", *Metrologia* **43** (1), 89-97, 2006.