Evaluating methods of calculating measurement uncertainty

B. D. Hall
Measurement Standards Laboratory of New Zealand,
Industrial Research Ltd, PO Box 31-310, Lower Hutt, New Zealand.
(email: b.hall@irl.cri.nz)

Abstract

This communication demonstrates the need for independent validation when an uncertainty calculation procedure is applied to a particular type of measurement problem. A simple measurement scenario is used to highlight differences in the performance of two general methods of uncertainty calculation, one from the *Guide to the Expression of Uncertainty in Measurement* (GUM) and one from *Supplement 1 to the "Guide to the Expression of Uncertainty in Measurement" – Propagation of Distributions using a Monte Carlo method*. The performance of these methods is investigated in terms of the long-run success-rate when applied to many simulated measurements in the scenario. An individual application of the method is deemed successful if an uncertainty interval containing the measurand is obtained. The alternative approach to validation taken in the *Supplement*, that an uncertainty interval calculated by a Monte Carlo method can be used to validate the GUM method, is not consistent with the results of this study.

1 Introduction

The analysis of a particular set of experimental data by several different uncertainty calculation procedures may yield different results for the expanded uncertainty interval at a given coverage level. In such cases, it would be very helpful to have recourse to a method of evaluating the performance of the different procedures.

The long-run success-rate of an uncertainty calculation procedure can be taken as a performance measure for such purposes. For example, a procedure claiming a 95% coverage level should (when applied to a large number of independent measurements) calculate expanded uncertainty intervals that contain the measurand on close to 95% of occasions. This seems a very reasonable requirement for any uncertainty calculation procedure. Any evidence of a much lower success-rate surely compromises the procedure's usefulness for evaluating uncertainty: why report a nominal 95% measurement uncertainty if the procedure's success-rate is significantly less than that?

This communication uses a simple measurement scenario to highlight a difference in performance between two general methods of uncertainty calculation and therefore a need for independent assessment. The first method is described in the *Guide to the Expression of Uncertainty in measurement* (GUM) [1] and the second is described in the first Supplement to the GUM, which is soon to be published (SUP) [2]. In the scenario, the measurand is a non-linear function of two independent input quantities. Non-linearity can be difficult for uncertainty-calculation procedures to handle, so an investigation of the performance of the two methods is sensible. In particular, the GUM

method is not expected to perform well with a non-linear measurement function, whereas the SUP method has been recommended in such cases. In the following sections, the long-run success-rate of each method is evaluated for a selection of different measurands by processing data from many simulated measurements.

2 The measurement scenario

We consider a measurement of the magnitude of a complex-valued quantity $\mathbf{\Gamma} = \Gamma_1 + \mathrm{i}\Gamma_2$. Measurements of the real and imaginary components yield $\mathbf{z} = z_1 + \mathrm{i}z_2$, from which an estimate of $|\mathbf{\Gamma}|$ can be found

$$|\mathbf{z}| = \sqrt{z_1^2 + z_2^2}\,. \tag{1}$$

We further consider that z_1 and z_2 are independent and have standard uncertainties equal to a known value u associated with a Gaussian distribution.

2.1 The GUM uncertainty calculation

Applying the *Law of Propagation of Uncertainty* to equation (1) [1], we first evaluate the partial derivatives

$$\frac{\partial|\mathbf{z}|}{\partial z_1} = \frac{z_1}{|\mathbf{z}|} \tag{2}$$

$$\frac{\partial|\mathbf{z}|}{\partial z_2} = \frac{z_2}{|\mathbf{z}|} \tag{3}$$

and then obtain the combined standard uncertainty

$$u^2(|\mathbf{z}|) = \left(\frac{\partial|\mathbf{z}|}{\partial z_1}\right)^2 u^2 + \left(\frac{\partial|\mathbf{z}|}{\partial z_2}\right)^2 u^2 \tag{4}$$

$$= \frac{z_1^2 + z_2^2}{|\mathbf{z}|^2}\, u^2 \tag{5}$$

$$= u^2\,. \tag{6}$$

So, in this case, the uncertainty obtained by the method adopted in [1] is not a function of the estimates z_1 and z_2 obtained by measurement. For a particular result, an uncertainty interval for $|\mathbf{\Gamma}|$ with a 95% coverage level is

$$[\,|\mathbf{z}| - 1.96u, |\mathbf{z}| + 1.96u\,]\,, \tag{7}$$

where $|\mathbf{z}|$ is calculated from (1).[1]

2.2 The SUP uncertainty calculation

The first Supplement to the GUM describes a Monte Carlo method of calculating uncertainty [2]. The application of this method here can be described in a series of steps.

[1] Some readers may be concerned that the lower bound calculated by this method can be negative, which is an impossible value for the magnitude of a complex number. Firstly, note that negative values do not pose a problem in evaluating the long-run success-rate, as described below. Secondly, if the lower bound of any interval is negative, it is a trivial additional step to set it to zero. Doing so is a common-sense use of the available information.

1. Generate a sequence of samples z_{1i}, where $i = 1, \cdots, L$ and L is a large number, by drawing from a Gaussian distribution with mean z_1 and variance u^2.
2. Generate corresponding sequence of samples z_{2i} by drawing from a Gaussian distribution with mean z_2 and variance u^2.
3. Calculate
$$|\mathbf{z}_i| = \sqrt{z_{1i}^2 + z_{2i}^2}$$
for $i = 1, \cdots, L$.
4. Sort the values of $|\mathbf{z}_i|$ in ascending order.
5. Take the $0.025L^{\text{th}}$ smallest value[2] of $|\mathbf{z}_i|$ as the lower bound of the uncertainty interval.
6. Take the $0.975L^{\text{th}}$ smallest value[3] of $|\mathbf{z}_i|$ as the upper bound of the uncertainty interval.

The result is an interval said to have 95% probability of containing $|\mathbf{\Gamma}|$.

2.3 The evaluation method

In order to assess the long-run success rates of the GUM and SUP methods, a series of simulated measurement results is processed using both methods. That is, pairs of data $(z_1[j], z_2[j])$, $j = 1, \cdots, 1000$, are simulated and each pair is used as if it were the data obtained from an independent measurement of the same fixed measurand. The procedure is as follows.

1. A value for $\mathbf{\Gamma} = \Gamma_1 + \mathrm{i}\Gamma_2$ is selected.
2. A sequence of pairs, $(z_1[j], z_2[j])$, is drawn from independent Gaussian distributions with means Γ_1 and Γ_2, respectively, and variances u^2.
3. For each pair, a 95% uncertainty interval is calculated using the GUM method and a counter is incremented if that interval contains $|\mathbf{\Gamma}|$.
4. For each pair, a 95% uncertainty interval is calculated using the SUP method and a counter is incremented if that interval contains $|\mathbf{\Gamma}|$.
5. The success rate of a procedure is estimated from the respective counter value.

These steps assess the success-rate of a procedure at one fixed value of the measurand. In order to investigate a procedure's performance over a range of values, the method can be repeated with different measurand values at step 1. This is the approach taken in the next section.

2.4 Numerical application

The evaluation method above was used to investigate the performance of the GUM and SUP procedures. The symmetry of the scenario means that performance will be independent of the radial (azimuthal) coordinate of the measurand in the complex plane. So, without loss of generality, seven $\mathbf{\Gamma}/u$ values lying along the real axis were chosen.

[2] If $0.025L$ is not an integer then take the largest integer that is less than $0.025L$
[3] If $0.975L$ is not an integer then take the smallest integer that is greater than $0.975L$

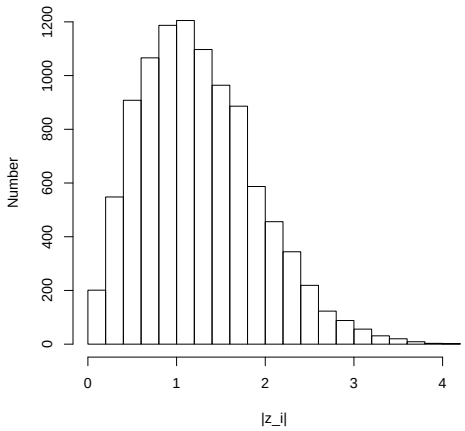

Figure 1: A histogram of magnitudes $|\mathbf{z}_i|$ calculated from from a Monte Carlo sample of 10000 generated with $z_1 = 0$, $z_2 = 0$ and $u = 1$. The $0.025L^{\text{th}}$ smallest magnitude in this sample is 0.224.

For each measurand value, simulated experiments provided 1000 sets of input data for the two uncertainty-calculation procedures.

The results are summarised in Table 1. The second and third columns of this table report the number of successes out of 1000 for the respective methods. In carrying out the SUP method, a Monte Carlo sample size of $L = 10^4$ was used.

$\|\mathbf{\Gamma}\|/u$	GUM	SUP
0.1	884	0
0.2	880	0
0.5	920	735
1.0	955	908
2.0	956	934
5.0	948	945
10.0	943	943

Table 1: The number of successes out of 1000 for the GUM and SUP methods for several choices of measurand at different distances from the origin.

A dramatic fall in the success-rate of the SUP method occurs when the measurand is close to the origin. The method fails to reach the required rates of success for $|\mathbf{\Gamma}|/u < 2.0$ and fails on every occasion when $|\mathbf{\Gamma}|/u < 0.2$.[4] The origin of this problem is illustrated in Figure 1, which shows a histogram of magnitudes $|\mathbf{z}_i|$ for a Monte Carlo data set generated with $z_1 = 0$ and $z_2 = 0$ and a standard uncertainty $u = 1$. The lower bound of the 95% uncertainty interval for this data is $|\mathbf{z}_i| = 0.224$, and we would expect this bound to increase for any other Monte Carlo sample, so any measurand $|\mathbf{\Gamma}| \leq 0.224$ will fall outside the uncertainty intervals that can be generated. This explains the success-rate of zero in the first two rows of Table 1.

[4]Some variability in the number of successes observed can be expected. For a success-rate of $p = 0.95$, the standard deviation of the number of successes observed is roughly $\sqrt{Np(1-p)}$, which is approximately equal to 7 with $N = 1000$.

3 Discussion

With the importance of traceability in metrology, methods used to calculate uncertainty should perform well in an event-based paradigm, because it is ultimately the accuracy of measurement and calibration events that is required. Failure of a method to do so is surely of concern. Consequently, valid methods of uncertainty calculation must achieve acceptable rates of success in the intended measurement scenarios.

Long run success-rate is an important performance measure for uncertainty calculation procedures. Some authors already take this view when discussing methods of uncertainty calculation (e.g.: [3–5]). Using success-rate as a performance measure is reminiscent of the use of level-of-confidence in frequentist statistics. However, adopting it as a performance requirement will not preclude other methods of uncertainty calculation. A recent discussion on the interplay of Bayesian and frequentist analysis suggested that a 'practical' Bayesian approach would also be inclined to accept such a requirement [6]. In addition, Willink has shown that success-rate would provide a clear interpretation of the meaning of an expanded uncertainty interval [7], because it does not appeal to the notion of probability, which differs between different schools of thought in statistics.

An alternative view on the validity of uncertainty calculation procedures seems to be widely held: that there is in some sense a 'correct' uncertainty interval for a measurement uncertainty problem in the absence of any concept of long-run behaviour. If this view is acceptable (see [8, section 2.2]), and if some procedure is known to generate this result, it can be used as a reference to validate other procedures (such as the GUM method). This view underlies the approach taken in the SUP, where the GUM method is considered invalid if a significant difference is observed between the endpoints of uncertainty intervals calculated by the GUM and SUP methods [2, section 8]. If that guideline is applied to the scenario used here, the SUP method will be retained and the GUM method rejected. However, suppose, for example, that a traveling standard with a small value of $|\Gamma|$ is circulated among the participants of measurement comparison similar to this scenario. Would it be wise to take uncertainty statements obtained by the SUP or GUM methods at face-value? What meaning should be given to a nominal 95% uncertainty interval that is *unlikely* to contain the value of the quantity intended to be measured?

Although the SUP method has not performed well in this study, it is a pragmatic general-purpose procedure that can achieve good long-run behaviour [6]. It is not intended here to suggest otherwise. We expect that examples could be found in which the SUP method performs better than the GUM method. However, there are certainly other scenarios in which the long-run success-rate of the GUM method will be superior. The point is that there are no guarantees, so an independent method of validation is needed.

An additional performance measure is needed when different procedures achieve success-rates at or above their nominal coverage levels. A procedure that achieves a high success-rate may be generating wider intervals, on average, than one with a success-rate closer to nominal. So a secondary concern is a procedure's ability to produce narrow intervals, which represent stronger measurement statements (see [7, 2.3]). This idea has been used to compare procedures to evaluate the two-dimensional uncertainty region for an estimate of a complex quantity [4].

It might be argued in relation to this study that the SUP describes several recipes for constructing an expanded uncertainty interval from a Monte Carlo sample. Indeed, after noting the problems above, a sensible course of action would be to consider whether refinements to a method used could improve its performance in this scenario.[5] However, the fundamental question cannot be avoided: how should the validity such refined procedures be determined, if not by some independent measure of long-run success?

Acknowledgment

The author is grateful to R. Willink for encouragement in preparing this communication and to D. R. White for helpful comments. This work was funded by the New Zealand Government as part of a contract for the provision of national measurement standards.

References

[1] BIPM IEC IFCC ISO IUPAC IUPAP and OIML 1995 *Guide to the Expression of Uncertainty in Measurement* 2nd edn (Geneva: International Organisation for Standardization)

[2] JCGM 2006 *Supplement 1 to the "Guide to the Expression of Uncertainty in Measurement" – Propagation of Distributions using a Monte Carlo method* (Final Draft)

[3] Wang C M and Iyer H K 2005 Propagation of uncertainties in measurements using generalized inference *Metrologia* **42** 145–153

[4] Hall B D 2006 Monte Carlo uncertainty calculations with small-sample estimates of complex quantities *Metrologia* **43** 220-226

[5] Willink R 2007 A generalization of the WelchSatterthwaite formula for use with correlated uncertainty components *Metrologia* **44** 340–349

[6] Bayarri M J and Berger J O 2004 The Interplay of Bayesian and Frequentist Analysis, *Statistical Science* **19**(1) 58–81

[7] Willink R 2006 Principles of probability and statistics for metrology *Metrologia* **43** S211-S219

[8] Willink R 2006 On using the MC method to calculate uncertainty intervals *Metrologia* **43** L39–L42

[5] For example, as an alternative to steps 5 and 6 of the procedure described in section 2.2, the shortest interval that contains $0.095L$ of the Monte Carlo sample of $|\mathbf{z}_i|$ values can be used [2, 7.6.2]. However, although it may perform better than the SUP procedure used here, the success-rates for this procedure too must inevitably fall to zero when a measurand is sufficiently close to the origin.