NPL
National Physical Laboratory

NPL REPORT MS 27

# MEANINGFUL EXPRESSION OF UNCERTAINTY IN MEASUREMENT

MAURICE COX
ANTHONY O'HAGAN

MARCH 2021

# Meaningful expression of uncertainty in measurement

Maurice Cox
Data Science Department, National Physical Laboratory, UK

Anthony O'Hagan
School of Mathematics and Statistics, University of Sheffield, UK

ISSN 1754-2960

https://dio.org/10.47120/npl.MS27

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW



Approved on behalf of NPLML by
Dr P M Harris, Science Area Leader for Data Analytics Modelling.

# EXECUTIVE SUMMARY

The Guide to the expression of uncertainty in measurement (GUM) has been the enduring guide on measurement uncertainty for metrologists since its first edition in 1993. According to the GUM, a measurement should always be accompanied by a reasoned and defensible expression of uncertainty, and the primary such expression is the standard uncertainty. In this article, we distinguish between the use of an expression of uncertainty as information for the recipient of a measurement result, and its use when propagating uncertainty about inputs to a measurement model in order to derive the uncertainty in a measurand. We propose a new measure of uncertainty, the *characteristic uncertainty*, and argue that it is more fit for these purposes than standard uncertainty.

For the purpose of reporting a measurement result, we demonstrate that standard uncertainty does not have a meaningful interpretation for the recipient of a measurement result and can be infinite. These deficiencies are resolved by the characteristic uncertainty, which we therefore recommend for use in reporting. For similar reasons, we advocate the use of the median estimate as the measured value.

For the purpose of propagating uncertainty in a measurement model, we propose simple propagation of the median and characteristic uncertainty and show through some examples that this *characteristic uncertainty framework* is simpler and at least as reliable and accurate as the propagation of estimate, standard uncertainty and effective degrees of freedom according to the GUM uncertainty framework.

# CONTENTS

# TABLES

# 1 INTRODUCTION

A basic premise of the Guide to the expression of uncertainty in measurement (GUM) [3] is that many measurements are modelled by a functional relationship, termed the *measurement model*, between $N$ input quantities $X_1,\ldots,X_N$ and an output quantity (or measurand) $Y$ in the form

$$Y = f(X_1,\ldots,X_N).$$

The *measurement function* $f$ may be mathematical or algorithmic.

The guiding principle of the GUM is that a measurement should always be accompanied by a reasoned and defensible expression of uncertainty. The GUM provides simple procedures for expressing uncertainty in the input quantities and, given these, to derive the uncertainty in the measurand. Thus the *measurement result* comprises the estimate, or *measured value*, of the measurand, together with an expression of uncertainty.

The GUM identifies two ways to evaluate the uncertainty in an input, which it names Type A and Type B evaluation. Type A evaluation for a quantity involves statistical analysis of data. Typically, the data will consist of a sample of individual estimates of the quantity, often referred to as *indications*, which are subject to random observation errors. A Type B evaluation is a judgement based upon the metrologist's expertise, published information, etc.

To derive an estimate and standard uncertainty of the measurand, the GUM advocates the use of the Law of Propagation of Uncertainty (LPU), but recognises that this is only applicable when the measurement model is linear or nearly linear.

Limitations in the scope of the GUM have been addressed in two supplements. Supplement 1 (GUM-S1) [7] provides a methodology to compute uncertainty in the measurand for more complex measurement models, and Supplement 2 [4] treats multivariate output quantities.

Despite its status as a key document for metrologists that is used in thousands of calibration and testing laboratories around the world, the GUM and its supplements have attracted sustained criticism and debate. Much of this has centred on fundamental and philosophical differences between Type A and Type B evaluations and the nature of the expressions of uncertainty that arise from them.

## 1.1 TERMINOLOGY

Before proceeding further, we will clarify our use of the terms 'measurement', 'measurand' and 'measurement result'.

The VIM [6, clause 2.1] defines measurement to be a 'process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity', but this is a vague and ambiguous definition. The word 'experimentally' seems to limit measurement to a process that is conducted as an experiment. As such, it would seem to encompass Type A evaluation of a quantity, if the data employed for the evaluation have been obtained 'experimentally', but to exclude Type B evaluation. And since a measurement model often combines inputs subject to both kinds of evaluation, the use of such a model is also not measurement according to this definition. Needless to say, metrologists always regard the use of a measurement model as measurement, and therefore this definition does not accord with practice in metrology.

Measurement should certainly be a process. We believe that Type A evaluation, Type B evaluation and application of a measurement model are all processes that can and should be deemed to be measurements.

The VIM [6, clause 2.3] further defines a measurand to be a 'quantity intended to be measured'. We therefore regard any quantity that is the subject of Type A evaluation, Type B evaluation or application of a measurement model to be a measurand. In a Type A evaluation of a quantity $X$, $X$ is the measurand. When a measurement model expresses a quantity $Y$ in terms of other quantities $X_1,\ldots,X_N$, then $Y$ is the measurand in that measurement, and although the $X_i$ are the measurands in each of their respective evaluations, in this context they are referred to simply as inputs to the measurement of $Y$.

The VIM [6, clause 2.9] goes on to define the measurement result as a 'set of quantity values being attributed to a measurand together with any other available relevant information'. This definition deliberately allows a very wide range of interpretations, but the essence is that a measurement result should express, in some usable form, what is known about the measurand following its measurement. This is the interpretation that we will employ.

A measurement result for a quantity $X$ has two primary functions. The first is to inform a person who is interested in the value of $X$, and who receives the measurement result as information. We refer to this person as a *recipient* of the measurement result. The second function arises when $X$ is an input to a measurement model for quantity $Y$. We refer to this as the measurement result being *transferred* to the measurement of $Y$. Therefore, when we say that 'a measurement result should express, *in some usable form*, what is known' about $X$, it is important that is is usable for both functions.

### 1.2 Overview

The paper is organised as follows. An expression of the uncertainty concerning a measurand is generally seen as an essential part of a measurement result. Although the GUM introduces the standard uncertainty as the primary expression of measurement uncertainty, Section 2 identifies a number of ways in which it may be both problematic and unhelpful for the recipient of a measurement result.

Section 3 proposes an alternative measure of uncertainty. We would argue that this measure has at least an equal case to be called 'standard uncertainty', except that that term is already confusingly used for several different kinds of standard deviation. Our new measure is therefore referred to herein as *characteristic uncertainty*. Characteristic uncertainty resolves the problems associated with standard uncertainties and is more meaningful and readily interpretable by recipients of measurement results.

Section 3 goes on to consider the term 'measured value', whose definition in the GUM and the VIM is also ambiguous. The median value is proposed as a more meaningful measured value and this section concludes with a wider discussion of meaningful ways to report the result of a measurement, for the benefit of a recipient of that measurement.

In Section 4 we turn to the second function of a measurement result. The result of a measurement of a quantity $X$ must be not only meaningful but also transferable, i.e. usable to compute the measurement result for $Y$ when $X$ is an input to the measurement of $Y$. One advantage that is claimed for the standard uncertainty is that it is transferable using the LPU, which gives the standard uncertainty of $Y$ exactly in the case of a linear measurement model and approximately for a model that is 'nearly linear'. Section 4 considers propagation and transferability for both standard uncertainties and characteristic uncertainties, concluding from some examples that our proposed reporting measures of median and characteristic uncertainty have at least equally good transferability properties.

Section 5 summarises the key conclusions, that our new expressions of uncertainty, namely the median value and the characteristic uncertainty, are not only more meaningful than the usual estimates and standard uncertainties for reporting the result of a measurement but also have at least equally good properties when propagating uncertainty through a measurement model. We also acknowledge the limitations of any simple two-number summary, and emphasise that ultimately it is the full probability distribution of a measurand that must be the primary result of a measurement.

## 2 STANDARD UNCERTAINTY

The GUM introduced the *standard uncertainty*, which has been universally adopted in metrology as the primary expression of uncertainty in measurement. The International Vocabulary of Metrology (VIM) [6] defines standard uncertainty to be a standard deviation. However, this definition has always been ambiguous because standard uncertainties can be derived in several distinct ways, with quite different interpretations.

### 2.1 FREQUENCY STANDARD UNCERTAINTY

The first of these standard uncertainties arises in the GUM in Type A evaluation of uncertainty, where an estimate of a quantity has been obtained by statistical analysis of some data. If an estimate $x$ for a quantity $X$ has been obtained in this way, the GUM defines the standard uncertainty to be (an estimate of) the standard deviation of the estimator.

We will refer to a standard uncertainty of this type as a *frequency standard uncertainty* and denote it by the symbol $u_{\mathrm{f}}(x)$, because the statistical methodology assumed by the GUM for Type A evaluations is the frequentist paradigm. The frequency probability for an event is defined as the long run rate at which that event would occur in an infinitely long sequence of instances in each of which that event may or may not occur. In frequentist theory, all probabilities are frequency probabilities. Thus, $u_{\mathrm{f}}(x)$ only has meaning in the context of a hypothetically infinite sequence of samples of data. If the estimate $x$ is computed for each of these samples, then $u_{\mathrm{f}}(x)$ is (an estimate of) the standard deviation of these values.

In frequentist theory, probabilities cannot be assigned to $X$, because it is not random or repeatable. A quantity in a measurement model, whether it be the measurand itself or an input quantity, has a fixed, unknown value for the measurement at hand. It is not random and cannot have frequency probabilities. Although a frequency standard uncertainty is typically interpreted as a description of uncertainty about the quantity $X$, strictly it is a measure of variability of the estimate $x$. Hence the argument of $u_{\mathrm{f}}$ is $x$.

### 2.2 BAYESIAN STANDARD UNCERTAINTY

GUM Supplement 1 (GUM-S1) [7] derives standard uncertainty in a different way for Type A evaluations. It uses the Bayesian statistical paradigm to analyse the data.

Bayesian theory adopts a different definition of probability, known as *personal probability*, or subjective probability. Instead of the frequency definition, probability is an expression of personal belief, experience and knowledge. Personal probability can apply to any uncertain quantity or event, without a requirement for repeatability. In particular, the fixed quantity $X$ has a probability distribution that represents what is known about it. Bayesian analysis distinguishes between the *prior* distribution of $X$, which represents what is known about $X$ before seeing the data, and its *posterior* distribution, which represents what is known after seeing

the data. Bayes' theorem is applied to combine the prior distribution with the data to yield the posterior distribution.

The Bayesian standard uncertainty $u_b(X)$ is the standard deviation of the posterior distribution of $X$, and is therefore a direct expression of uncertainty about $X$, in the light of the observed $x$. The argument of $u_b$ is therefore $X$.

### 2.3 JUDGEMENT STANDARD UNCERTAINTY

In the GUM, Type B evaluation of uncertainty is not derived from analysis of data. Instead, the standard uncertainty is a judgement by the metrologist of the quality of the metrologist's estimate $x$ for $X$. We will refer to this as a *judgement standard uncertainty* and denote it by $u_j(X)$.

A judgement standard uncertainty implicitly uses personal probability, and differs only from a Bayesian standard uncertainty by being expressed directly by the metrologist, rather than being derived from a Bayesian analysis of data. Nevertheless, it should be formulated in the light of all the available knowledge and expertise. As with Bayesian evaluation, the argument of $u_j$ is $X$.

### 2.4 COMBINED UNCERTAINTY

The GUM asserts that where a measurement model expresses a measurand in terms of some inputs with frequency standard uncertainties and some with judgement standard uncertainties, they can be combined in linear or near-linear models using the law of propagation of uncertainty (LPU) to yield the standard uncertainty of the measurand. Strictly, these disparate forms of standard uncertainty cannot be combined in that way. They certainly cannot be combined in frequency probability terms, because the subjective standard uncertainties cannot have any meaning in frequency terms. The GUM claims that the LPU is nevertheless legitimate, but offers conflicting justifications for this assertion. In [3, clause G.4.2] it implies that a judgement standard uncertainty $u_j$ can be treated as a frequency standard uncertainty, and offers a way to assign a degrees of freedom based on a 'relative uncertainty' in the metrologist's judgement of $u_j$. Personal probability does not recognise such an 'uncertainty about uncertainty' and the GUM does not indicate how the metrologist can contemplate such a thing. Even if a value can be obtained for a degrees of freedom in this way, frequency standard uncertainties can only be defined in relation to repeated realisations of a random process. We find this proposed justification for combining frequency and judgement standard uncertainties wholly unconvincing.

More credible is the contrary suggestion in [3, clause E.3.5] that, disparate standard uncertainties can be combined because ultimately all expressions of uncertainty must be the metrologist's judgement and opinion, even when based on Type A evaluation of uncertainty, and judgement uncertainties can be legitimately combined. This attitude would indeed be convincing if the frequency standard uncertainties can be viewed as judgements using personal probability. However, the GUM does not explain the mechanism by which a Type A frequency standard deviation $u_f(x)$, a property of the estimate $x$ defined over hypothetical repeated sampling, becomes a judgement of uncertainty $u_j(X)$ about the quantity $X$ in the sense of personal probability.

The Bayesian approach of GUM-S1 offers a resolution of this disparity. Both Bayesian and judgement standard uncertainties are based on personal probability judgements, and the standard uncertainties from all inputs to a measurement model can then legitimately be combined

to obtain a standard uncertainty for the measurand in the personal probability sense. The combination can be through the LPU in the case of linear models, or through the Monte Carlo method advocated in GUM-S1 for measurement models with appreciable nonlinearity.

However, the approach in GUM-S1 yields a numerically different standard uncertainty from that in the GUM in some typical measurement problems.

Consider the canonical Type A evaluation where the data comprise $n$ independently obtained indications $x_1,\ldots,x_n$ having the normal distribution with unknown mean $\mu$ and unknown variance $\sigma^2$. The best estimator of $\mu$ is the sample mean $\bar{x}$. In the GUM, the frequency standard uncertainty is given as $u_{\mathrm{f}}(\bar{x}) = s/\sqrt{n}$, where $s^2$ is the experimental variance $\sum(x_i-\bar{x})^2/(n-1)$.

For this problem, GUM-S1 applies a standard Bayesian analysis based on an uninformative prior distribution and obtains the Bayesian standard uncertainty $u_{\mathrm{b}}(\mu) = \sqrt{(n-1)/(n-3)}s/\sqrt{n}$, which is larger than the GUM's $u_{\mathrm{f}}(\bar{x})$ by the factor $\sqrt{(n-1)/(n-3)}$. For small samples, the difference can be large, for instance the factor is $\sqrt{2}$ when $n = 5$, and $u_{\mathrm{b}}(\mu)$ does not exist (and is effectively infinite) when $n < 4$. This increase in standard uncertainty is viewed as unpalatable by many metrologists.

### 2.5 Interpreting standard deviation

The preceding discussion has highlighted some of the problems with defining the primary expression of uncertainty to be a standard uncertainty. First, there are different ways of constructing a standard uncertainty, with different philosophical underpinnings and leading to numerically different values even in the most basic of measurements.

Second, the standard uncertainty can be infinite, and to report that measurement uncertainty is infinite would not reflect well on the metrologist conducting the measurement. The situations in which this arises are not limited to Bayesian evaluations, as described in Appendix A. Undefined or infinite standard uncertainties are just one aspect of the underlying fact that the standard deviation of a probability distribution is highly sensitive to its tails. Tiny amounts of probability for extreme values of a quantity can substantially increase the standard deviation. Therefore, instead of expressing how far an estimate might typically deviate from the measurand's value, the standard uncertainty may be simply an artefact of the tail shape of the probability distribution.

These problems already cast doubt on the usefulness of a standard uncertainty to the recipient of a reported measurement result, which is the first of the two purposes of the measurement result identified in Section 1.1.

More importantly, from the recipient's perspective, what meaningful information does a standard uncertainty $u$ convey about a measurand $X$?

A recipient might typically think that $X$ will probably be within one standard uncertainty of the estimate, and that it is very likely (perhaps about 95 % certain) to be within two standard uncertainties of the estimate. These are vague interpretations of a standard uncertainty.

Furthermore, this usual interpretation of a standard deviation can be quite wrong, depending on the tail shape of the distribution. In the case of the single normal sample the probability that $X$ lies within two frequentist standard uncertainties of the estimate is much less than 95 % if the sample size is small. Recognising this difficulty of interpreting the standard uncertainty, the GUM defines the *expanded uncertainty* $U(\bar{x})$ to be such that the interval $\bar{x} \pm U(\bar{x})$ has 95 % coverage. For small sample sizes, $U(\bar{x})$ is appreciably larger than $2u_{\mathrm{f}}(\bar{x})$.

We conclude that no concrete, quantitative, meaningful interpretation of a standard uncertainty is possible.

# 3 MEANINGFUL REPORTING OF MEASUREMENT

In this section, we will consider more meaningful ways to report the result of measurement.

We will adopt the Bayesian paradigm in what follows, because we are convinced that it is the only sound and logically consistent framework for metrology, when Type A and Type B evaluations must be combined coherently, and when a metrologist must necessarily use judgement and expertise at all stages of a measurement. In taking this position, we follow [3, 12, 14, 17, 18].

From the Bayesian perspective, uncertainty in any quantity is expressed using probabilities. In particular, a complete description of uncertainty in a measurand consists of a probability distribution (or PDF). In the case of a Bayesian Type A evaluation this is the posterior distribution; for a Type B evaluation it can be expressed as the metrologist's judgement. Distributions for the inputs to a measurement model imply a probability distribution for the measurand, which may for instance be computed using the Monte Carlo method of GUM-S1. Probabilities and probability distributions are always to be understood as representing the considered opinion and judgement of the metrologist.

## 3.1 CHARACTERISTIC UNCERTAINTY

For a measurand $X$ with estimate $m(X)$, we define the *characteristic uncertainty* of $X$, denoted by $c(X)$, to be such that $m(X) \pm 2c(X)$ is a 95 % coverage interval for $X$.

Unlike a standard uncertainty, defined as a standard deviation, a characteristic uncertainty has a clear and meaningful interpretation for the user of a measurement result. It always exists and conveys concrete information about $X$. Instead of the vague interpretations that are typically (and sometimes erroneously) attributed to a standard deviation, the interpretation of a characteristic uncertainty is that $X$ lies within $2c(X)$ of the estimate $m(X)$ with probability 95 %, no more, no less. As with standard uncertainty, the user can also expect that $X$ will probably lie within one characteristic uncertainty of the estimate.

We believe that the characteristic uncertainty should form the principal expression of uncertainty when reporting a measurement result, on the grounds that it is more useful and meaningful to the recipient than a standard uncertainty.

We note that in many metrology applications, a 95 % coverage interval is usually specified as part of a measurement result, typically by specifying the expanded uncertainty, and it may even be given more prominence than the standard uncertainty. The characteristic uncertainty conveys essentially the same information as this when reporting a measurement result, but we believe that recipients of such reporting will benefit from being consistently given this clear and meaningful expression of measurement uncertainty. Furthermore, we will show in Section 4 that its value extends also to when $X$ becomes an input to another measurement model, which is the second purpose of a measurement result.

## 3.2 THE NORMAL SAMPLE CASE

To illustrate the value of this new uncertainty measure, we consider characteristic uncertainty in the canonical Type A evaluation context of a normal sample, as described in Subsection 2.4.

Although the frequentist and Bayesian methods lead to different standard uncertainties, they both give the same estimate $m(X) = \bar{x}$ and 95 % coverage interval for $X$: $\bar{x} \pm k_{n-1} s/\sqrt{n}$, where $k_d$ is the upper 97.5 % point of the Student t distribution with $d$ degrees of freedom. Therefore the characteristic uncertainty is

$$c(X) = k_{n-1} s/(2\sqrt{n}),$$

regardless of whether the metrologist employs the frequentist statistical paradigm of the GUM or the Bayesian paradigm of GUM-S1. Characteristic uncertainty thereby resolves, in this most basic and widely used analysis in metrology, the conflict between the frequency and Bayesian standard uncertainties.

Table 1 gives values of $c(X)/u_{\mathrm{f}}(\bar{x}) = k_{n-1}/2$ for various values of the sample size $n$. These numbers are familiar as half the expanded uncertainty factor for the normal sample problem. For $n < 10$, this factor is appreciably larger than 1, and hence $c(X)$ is larger than $u_{\mathrm{f}}(\bar{x})$. This highlights the deficiency of the frequentist standard uncertainty as a meaningful expression of uncertainty. With small sample sizes, the simple notion that with probability about 95 % $X$ will be within two standard deviations of the estimate is seriously erroneous and optimistic.

**Table 1:** Comparisons between $c(X)$ and $u_{\mathrm{b}}(x)$ for the normal sample

| $n$ | $c(X)/u_{\mathrm{f}}(\bar{x})$ | $u_{\mathrm{b}}(x)/u_{\mathrm{f}}(\bar{x})$ |
|---|---|---|
| 2 | 6.35 | $\infty$ |
| 3 | 2.15 | $\infty$ |
| 4 | 1.59 | 1.73 |
| 5 | 1.39 | 1.41 |
| 7 | 1.22 | 1.22 |
| 10 | 1.13 | 1.13 |
| 20 | 1.05 | 1.06 |
| $\infty$ | 0.98 | 1.00 |

The final column of table 1 gives values of the factor $u_{\mathrm{b}}(x)/u_{\mathrm{f}}(\bar{x}) = \sqrt{(n-1)/(n-3)}$ and it is noticeable how close they are to the values of $c(X)/u_{\mathrm{f}}(\bar{x})$ in the second column for $n > 4$. Therefore, unless $n \leq 4$ the characteristic uncertainty is very close to the Bayesian standard uncertainty. Figure 1 shows the ratio $c(X)/u_{\mathrm{b}}(x)$ as a function of $n$ for $n$ up to 500. For $n > 7$ this ratio decreases smoothly with $n$, asymptotically approaching 0.979 98 ....

As explained earlier, we adopt the Bayesian perspective, and so we regard the Bayesian $u_{\mathrm{b}}(X)$ as the appropriate standard uncertainty in this problem. Table 1 shows that for sample sizes larger than 4, the recipient of a report containing this standard uncertainty would not be seriously wrong in understanding that the measurand is about 95 % certain to lie within two standard uncertainties of the estimate. Whereas the comparable intuition for the frequentist standard uncertainty would be substantially wrong unless $n$ is more than 10.

Sample sizes in practical metrology are very often smaller than 10, and may indeed be smaller than 4. The characteristic uncertainty by definition has the desired interpretation for all sample sizes of 2 or more, and is always finite.

### 3.3 The median estimate

Having proposed the characteristic uncertainty as a more useful and meaningful expression of uncertainty for the recipient of a measurement result, we now turn our attention to the

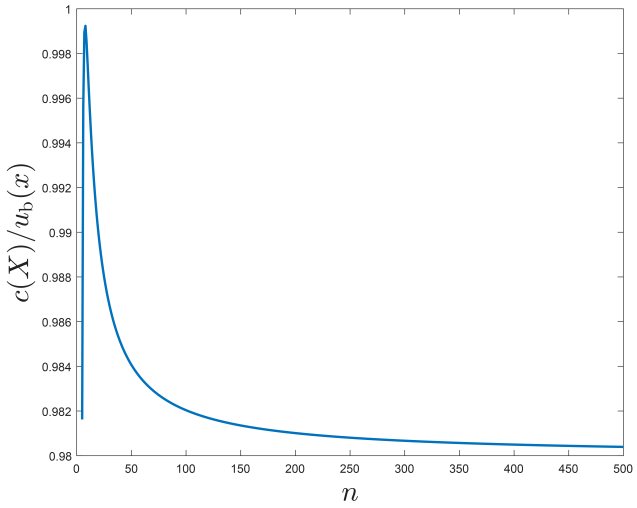


**Figure 1:** Ratio $c(X)/u_{\mathrm{b}}(x)$ as a function of sample size $n$

measured value. The notion of a measured value in metrology is even more ambiguous than the standard uncertainty. A measurement is a process that rarely consists simply of reading a single number from a physical instrument, so the term 'measured value' refers to a number deriving from that process that can be variously referred to as a 'representative value', an 'estimate', a 'best estimate' or an 'expected value'.

A frequentist Type A evaluation will typically result in an estimate, which may formally be an unbiased estimate.

The result of a Bayesian Type A evaluation or a Type B evaluation will be a probability distribution, and it is usual to choose the mean (also known as the expectation) of this distribution as the measured value.

When using a measurement model, the measured value according to the GUM uncertainty framework is simply the result of plugging 'measured values' of all the inputs into the measurement function. (We note an alternative suggestion in [3, clause 4.1.4] to average such values where replication is available.) If, however, the Monte Carlo method is used, it is specified to be the mean of the distribution of the measurand [7, clause 5.1.1].

As with the standard uncertainty, we ask what useful interpretation the recipient or user of a measurement result can place on the measured value. A 'representative' value can be arbitrary, at the whim of the metrologist. An 'estimate' could be the result of applying any estimation method, good or bad. The result of plugging measured values of inputs into a measurement model is just that, with no other formal interpretation. It is often referred to as a 'best estimate', but without justification or explanation of in what sense it is 'best'.

A measured value that is the mean of the measurand is at least well defined (when it exists; see Appendix A), but in practice it is not clear what that value would convey to the user. Where the metrologist's judgement about the measurand is represented by a symmetric probability distribution, as in the case we have been considering of Type A evaluation from a single normal sample, there is a natural best choice of a measured value — the mean or expected value lies at the centre of symmetry when it exists, and this is also the median and the mode (assuming the distribution is unimodal). However, although not treated explicitly in the GUM, asymmetric

distributions can arise in metrology and it is not so obvious that the mean is then a useful estimate.

Furthermore, as discussed in Appendix A, the mean may not exist.

We propose that the median of the measurand's probability distribution is a more useful and meaningful measured value. Compared with the mean, the median is typically located more in the central part of a skew distribution, where the probability density is highest; see Appendix C. More importantly, it always exists and has a clear and useful interpretation: the true value of the measurand is equally likely to be above or below the median.

The characteristic uncertainty $c(X)$ was defined in Subsection 3.1 by reference to the estimate $m(X)$, which we now formally identify as the median. Thus, we define $c(X)$ to be such that there is 95 % probability that $X$ will lie within $\pm 2c(X)$ of the median $m(X)$.

### 3.4 The new measures in practice

To show the practical implications of using the median and characteristic uncertainty we consider three examples. Formulae and methods for computing these new measures are given in Appendix B

**t distribution.** Our first example is a scaled and shifted Student t distribution, arising from a normal sample as discussed in previous sections. We suppose that the sample size is $n = 6$, the sample mean is $\bar{x} = 0$ and $s/\sqrt{n} = 0.0225$. Therefore the Bayesian or judgement distribution for $X$ is a t distribution with mean $\mathrm{E}(X) = 0$ and standard uncertainty $u(X) = 0.0225\sqrt{5/3} = 0.0290$. This distribution is shown in figure 2. Its median, $m(X)$, is also zero and the characteristic uncertainty is $c(X) = 0.0225\,k(5)/2 = 0.0289$.

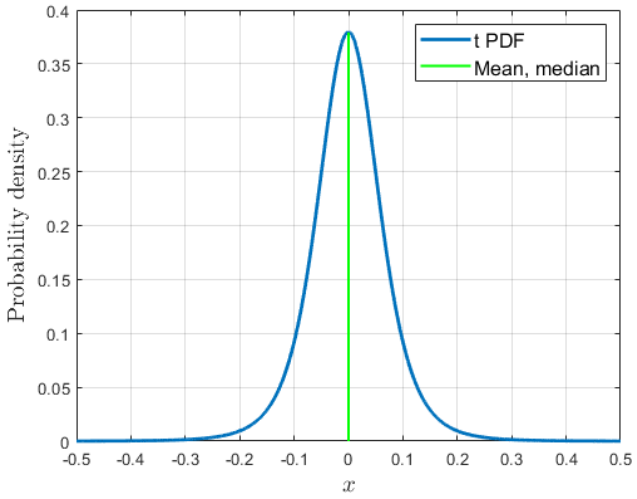


**Figure 2:** Density function of t distribution

**Gamma distribution.** A gamma distribution is appropriate for a quantity that must be positive, and can therefore arise in metrology as a Type B evaluation for such a quantity. Suppose that $X$ has the gamma distribution Ga(95, 7.6) with density shown in figure 3. It has mean $\mathrm{E}(X) = 0.0800$ and standard uncertainty $u(X) = 0.0290$. However, the median is $m(X) = 0.0765$ and the characteristic uncertainty is $c(X) = 0.0275$.

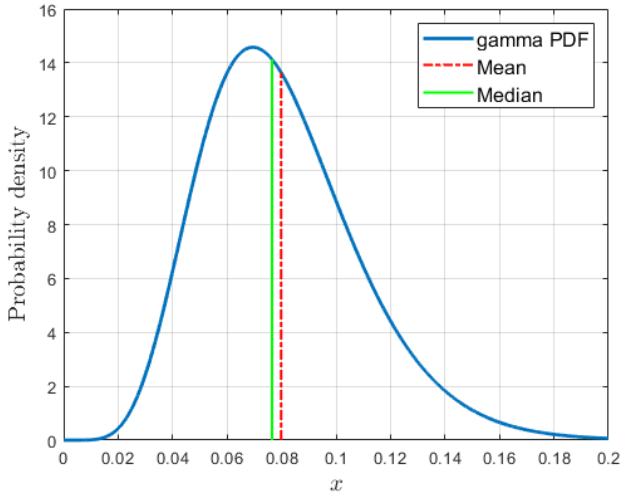


**Figure 3:** Density function of gamma distribution

**Skew-normal distribution.** The skew-normal family of distributions [2, 16] has a variety of applications in statistics. In metrology, it can arise when a measurand is the sum of two inputs, one of which has a constrained distribution (such as the half-normal distribution in Appendix C). Therefore suppose a measurand $X$ has the skew-normal distribution $\mathrm{SN}(-0.0355, 0.0458^2, 4)$ whose density is shown in figure 4. It has mean $\mathrm{E}(X) = 0$ and standard deviation $u(X) = 0.0290$. Its median, however, is $m(X) = -0.0046$ and its characteristic uncertainty is $c(X) = 0.0295$.

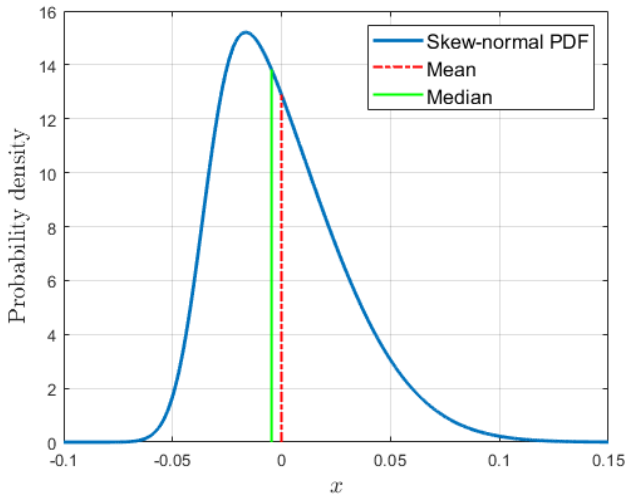


**Figure 4:** Density function of skew-normal distribution

In all three cases, the median is close to the mean and the characteristic uncertainty is close to the standard uncertainty. This will typically be the case in the majority of metrological contexts, and indeed the standard approaches set out in the GUM are designed for well-behaved problems like these, where the distribution of the measurand will be similar to a normal or t distribution. Therefore the new measures will not generally produce radically different values from the more familiar measures. However, we emphasise that the new median and characteristic uncertainty have clear, unambiguous and meaningful interpretations for the recipient.

They therefore fulfil the requirements of reporting a measurement result in ways that the traditional mean and standard uncertainty fail to do.

Situations can arise in metrology where the distribution of the measurand is not similar to the above examples, and exhibits considerable asymmetry. We discuss these briefly in Appendix C.

### 3.5 Reporting guidelines

We are now led to consider more widely the most useful and meaningful ways to report a measurement result, from the perspective of the recipient.

The result should represent the metrologist's considered judgement regarding the measurand, in the light of the available evidence and the metrologist's experience and expertise. We believe that the measurement result should always include the probability distribution that represents that judgement.

In the example of a sample from a normal distribution, it would be reported that $X$ has a scaled and shifted Student t distribution with median $m(X) = \bar{x}$ and characteristic uncertainty $c(X) = k_{n-1}s/(2\sqrt{n})$.

The probability distribution is a complete description of the metrologist's judgement regarding $X$. However, the distribution alone will not generally meet our requirements for reporting, because unless the recipient is well versed in statistics, it does not readily convey useful information about $X$. Therefore, various *summaries* of the distribution should be provided to convey clear and meaningful information for the recipient. It is for this purpose that we have developed the median and characteristic uncertainty. The median $m(X)$ is a summary measure of location, which can serve as an estimate or representative value of $X$. It has the specific meaning that $X$ is equally likely to be above or below $m(X)$. The characteristic uncertainty $c(X)$ is a summary measure of uncertainty, with the specific meaning that $X$ has a 95 % probability of lying within $m(X) \pm 2c(X)$.

Other summaries can usefully supplement these where appropriate, and we discuss such situations in Appendix C. However, we strongly advocate that the measurement result for a measurand $X$ should comprise the probability distribution for $X$ with at least the two new summary measures — the median and the characteristic uncertainty.

For all the reasons set out above, we see no reason to quote the standard uncertainty in reporting a measurement result. However, before rejecting it completely we must consider whether it still should be reported in case $X$ is subsequently to be used as an input to a measurement model for another measurand.

## 4 Propagation and transferability

In this section we will examine methods of propagating uncertainty through a measurement model, and their associated transferability properties.

We will refer to a group of components of a measurement result as *transferable* if there is a way to compute, at least approximately, those components for the measurand $Y$ given only those components for the model inputs $X_i$.

For instance, the mean and standard uncertainty comprise a transferable group if the measurement model is linear, because the mean and standard uncertainty of the measurand can be computed exactly from the means and standard uncertainties of the inputs using the LPU.

This transferability property is regarded as an important feature of the standard uncertainty, suggesting that the standard deviation should be included as a component of the measurement result for the second function described in Section 1.1.

However, we have argued that for reporting purposes the measurement result should include the median and characteristic uncertainty. Therefore, we shall only consider here exact or approximate methods that can deliver a measurement result for $Y$ that includes both the median of $Y$ and an expression of uncertainty that allows a 95 % coverage interval to be derived (e.g. the characteristic uncertainty of $Y$ or an expanded uncertainty).

Approximate propagation methods are widely used, and it should be noted that the approximation will be less accurate if the result components for the model inputs are themselves approximations (for instance, from being propagated through a sub-model [5]).

### 4.1 THE GUM UNCERTAINTY FRAMEWORK

The basic method advocated in the GUM has the following elements:

- The measurement model is assumed to express the measurand $Y$ as a linear function of the input quantities $(X_1,\ldots,X_N)$. If the model is not linear, then it is linearized about the estimates of the $X_i$ using the first-order Taylor series expansion.
- The estimate of $Y$ is obtained by plugging the estimates of the $X_i$ into the measurement function [3, clause 4.1.4]. For our purposes, this is taken to be an approximation to the median.
- The standard uncertainty of $Y$ is obtained or approximated by applying the LPU to combine the standard uncertainties of the $X_i$ in the (linearized) model.
- An 'effective' degrees of freedom $d$ for $Y$ is obtained by applying the Welch-Satterthwaite formula [15] to combine the standard uncertainties and degrees of freedom of the $X_i$ in the (linearized) model. The expanded uncertainty for $Y$ is then approximated as the standard uncertainty multiplied by $k_d$.

We will refer to this method as the GUM Uncertainty Framework (GUF) [7].

Notice that for the GUF it is the triplet of estimate, standard uncertainty *and* expanded uncertainty that is transferable. Estimates and standard uncertainties are propagated directly, while expanded uncertainties are propagated indirectly through the corresponding degrees of freedom. Degrees of freedom for the $X_i$ can be inferred from the ratio of their expanded and standard uncertainties, and then the expanded uncertainty for $Y$ is obtained from its standard uncertainty and its degrees of freedom.

The GUF is regarded as applicable in practice when it produces a sufficiently accurate measurement result for $Y$. The conditions for this to hold are generally argued as follows.

If the measurement model is nonlinear, then applying the GUF in the linearized model will give approximate values for the mean and standard uncertainty of $Y$. The approximation can be poor if the model is strongly nonlinear and the input standard uncertainties are large.

Furthermore, the LPU is only part of the GUF. The Welch-Satterthwaite formula is required to deliver the effective degrees of freedom, and hence the expanded uncertainty, but Welch-Satterthwaite is inherently approximate. Computing an expansion factor derived from Welch-Satterthwaite's effective degrees of freedom will only yield an approximate expanded uncertainty. The approximation is generally regarded as good if the distributions of the input quan-

tities are not too different from a normal or t distribution, and in particular if they are not markedly skew.

The GUF is therefore only considered to be applicable if the model is linear or nearly linear, and if the input distributions making substantial contributions to the uncertainty in $Y$ have a symmetric (or almost symmetric) form similar to a normal or t distribution [7, clauses 5.7, 5.8].

### 4.2 The Monte Carlo method

A primary objective of GUM-S1 was to overcome the limitation of the GUM uncertainty framework to linear or nearly linear models. For nonlinear models, GUM-S1 advocates a Monte Carlo method to compute the mean, standard uncertainty and a coverage interval for a stipulated coverage probability.

The GUM-S1 Monte Carlo method (MCM) requires more than the triplet of estimate, standard uncertainty and expanded uncertainty; the full probability distribution(s) of the $X_i$ must instead be specified. The method then has the following elements.

- Many random samples are drawn from the distributions of the $X_i$. For each sampled set of $X_i$ values, the measurement model is employed to provide a sampled value of $Y$.
- The resulting sample of $Y$ values represents the probability distribution of $Y$. The estimate $y$ and standard uncertainty of $Y$ are computed as the mean and standard deviation of the sample.
- Other summaries of this distribution may be readily computed, such as the median, characteristic uncertainty or a coverage interval for any stipulated coverage probability.

For the Monte Carlo method, it is the entire probability distribution that is transferable. Mean, median, standard uncertainty, expanded uncertainty, characteristic uncertainty or any other desired expressions of knowledge regarding the measurand are simply computed from the probability distribution: see Appendix B.1 for details of the computation of median and characteristic uncertainty.

From the Bayesian perspective Monte Carlo is the 'gold standard' and is always applicable because those expressions can be computed exactly (in the sense that they can be computed to any desired accuracy with a sufficiently large Monte Carlo sample). It is often the tool of choice in complex measurement problems such as those addressed in the National Metrology Institutes, but it is perceived by a large sector of the metrology community as technically and computationally more demanding than the GUM uncertainty framework.

The distribution of $Y$ must be reported, as recommended in Subsection 3.5, for transferability to be achieved; however, the Monte Carlo method delivers not the distribution itself but a large sample from it. One way to report the distribution is simply to provide the Monte Carlo sample. In a sense, this constitutes exact propagation, because if $Y$ then becomes an input to a second measurement model in which the Monte Carlo method is to be used, the reported sample is exactly what is needed in that second application of Monte Carlo.

Transferring a data set comprising a large sample of $Y$ values is entirely feasible with modern IT tools.

An alternative is to report a standard distribution fitted to that sample. If, for instance, the distribution is symmetric, unimodal and similar to a normal or t distribution, it can be reported

as the best-fitting such distribution. Whilst this may no longer represent exact propagation, a good approximation to the distribution of $Y$ will generally be adequate, and much simpler to report and transfer to a second measurement model than the full Monte Carlo sample.

See [11] for methods of obtaining a compact summary of the full Monte Carlo sample that preserves information about the measurand and can be used in a subsequent uncertainty evaluation.

### 4.3 THE CHARACTERISTIC UNCERTAINTY FRAMEWORK

We now suppose that we have medians and characteristic uncertainties for all input quantities in a measurement model, and consider how to propagate these in order to obtain the median and characteristic uncertainty for the measurand. Our simple proposal is to apply the same propagation rules as the GUF, but treating medians and characteristic uncertainties in the same way as means and standard uncertainties.

Our proposal therefore has the following elements:

- The measurement model is assumed to express the measurand $Y$ as a linear function of the input quantities $(X_1,\ldots,X_N)$. If the model is not linear, then it is linearized using the first-order Taylor series expansion about the medians.
- The median of $Y$ is approximated by plugging the medians of the $X_i$ into the measurement function.
- The characteristic uncertainty of $Y$ is approximated by applying the LPU to combine the characteristic uncertainties of the $X_i$ in the (linearized) model.

We will refer to this procedure as the characteristic uncertainty framework (CUF).

In the CUF it is the couplet of median and characteristic uncertainty that is transferable. It is therefore the simplest of the three propagation methods.

Whereas the GUF is exact when propagating means and standard uncertainties in linear measurement models, this is not true of the CUF. Even for a linear model the median and characteristic uncertainty of $Y$ given by the proposed propagation rules can only be approximate. Nevertheless, we argue that they will represent good approximations under the following conditions.

Provided the input distributions are not markedly skew, medians will be close to means, in which case plugging medians into the linear measurement function will yield a good approximation to the median of $Y$.

Furthermore, subsection 3.2 shows that any symmetric distribution that is close to a normal or t distribution with more than four degrees of freedom will have a characteristic uncertainty that is close to the corresponding standard deviation. Since the LPU is based on fundamental formulae for combining standard deviations, we can expect it to be a good approximation for characteristic uncertainties.

These intuitive arguments will be tested in Subsection 4.4.

We therefore propose that the CUF is applicable under the same conditions as the GUF, namely if the model is linear or nearly linear, and if the input distributions making substantial contributions to the uncertainty in $Y$ have a symmetric (or almost symmetric) form similar to a normal or t distribution.

### 4.4 COMPARISON

We will test the intuitive arguments we have given to suggest that the CUF should yield good approximations to the median and characteristic uncertainty of $Y$, by means of examples. In each case we will compare the median and characteristic uncertainty obtained in the characteristic uncertainty framework with (a) the gold standard values from Monte Carlo, and (b) the implied values given by the GUM uncertainty framework (the mean as approximation to the median and half the expanded uncertainty as approximation to the characteristic uncertainty). The Monte Carlo computations have been conducted with sufficiently large numbers of iterations to achieve accuracy to the stated numbers of significant figures.

EXAMPLE 1 *Two-term model*

A common measurement model takes the form

$$Y = X + C\,,$$

where the measurand $Y$ is modelled as a quantity $X$, evaluated as the sample mean of a set of $n$ normally distributed indications, plus an independent bias correction term $C$.

We suppose that the evaluation of $X$ is reported as a measured value of 5.7120 in some suitable units, with standard uncertainty $u(X)$. The expanded uncertainty for a 95 % coverage interval is reported as $u(X)k_{n-1}$. Under our proposal, it would simply be reported that $X$ has median 5.7120 and characteristic uncertainty

$$c(X) = u(X)k_{n-1}/2\,.$$

Our base case will be $n = 3$ and $u(X) = 0.0520$, while other cases will vary $n$ to 7 or $u(X)$ to 0.0260 or 0.0130. The case of $n = 3$ may seem extreme but it is common in routine metrology. Note that in this case the distribution of $X$ does not have a finite standard deviation, and so neither does $Y$. Their judgement standard uncertainties do not exist. Nevertheless, the characteristic uncertainty is well defined.

We will consider four different cases for the correction $C$. In each of these, $C$ is assigned a mean of 0 and a standard uncertainty of $u(C) = 0.0290$.

1. $C$ is evaluated by a Type B judgement. $C$ is assigned a normal distribution with mean (and median) 0 and standard deviation 0.0290. It therefore has characteristic uncertainty

   $$c(C) = 0.98 \times 0.0290 = 0.0284\,,$$

   where 0.98 is half of 1.96, the expanded uncertainty factor for a normal distribution. The normal distribution is defined to have infinite degrees of freedom.

2. $C$ is evaluated by a historic sampling exercise, together with the metrologist's judgement on how the bias in this instance might deviate from the historic data. $C$ is assigned a t distribution with 5 degrees of freedom, mean (and median) 0 and standard deviation 0.0290. Its characteristic uncertainty is therefore

   $$c(C) = 0.0145 \times k(5)\sqrt{3/5} = 0.0289\,.$$

3. $C$ is evaluated by a Type B judgement to the effect that the bias could be between −0.0502 and +0.0502. A uniform (rectangular) distribution is assigned between these bounds, which therefore has mean (and median) 0 and standard deviation 0.0290. The characteristic uncertainty is

   $$c(C) = 0.475 \times 0.0502 = 0.0238\,.$$

   By convention, the uniform distribution also has infinite degrees of freedom [9, section 2.5.4.1].

4. $C$ is evaluated by a Type B judgement reflecting the metrologist's opinion that $X$ is a little more likely to overestimate $Y$ than to underestimate. $C$ is assigned the skew-normal distribution of Subsection 3.3. It has mean 0, median $m(C) = -0.0046$, standard deviation 0.0290 and characteristic uncertainty $c(C) = 0.0295$. The tails of the skew-normal distribution are at least as thin as those of the normal distribution, and so this also has infinite degrees of freedom.

Combining four cases for the distribution of $X$ and four for the distribution of $C$, we have 16 cases in all. These are set out in the first four columns of table 2. For instance, the case denoted by 2.3 in the first column combines the second case of the distribution of $X$, in which the sample size is $n = 7$ and the standard uncertainty is $u(x) = 0.052$, with the third case of the distribution of $C$, which is the uniform distribution.

**Table 2:** Comparing GUM and characteristic uncertainty frameworks, Example 1

| Case | $n$ | $u(X)$ | $C$ | MCM $c(Y)$ | GUF $c(Y)$ | GUF % | CUF $c(Y)$ | CUF % |
|---|---|---|---|---|---|---|---|---|
| 1.1 | 3 | 0.052 | N | 0.1143 | 0.088 | 91.8 | 0.115 | 95.1 |
| 1.2 | 3 | 0.052 | $t_5$ | 0.1147 | 0.090 | 92.1 | 0.116 | 95.1 |
| 1.3 | 3 | 0.052 | U | 0.1141 | 0.088 | 91.8 | 0.114 | 95.0 |
| 1.4 | 3 | 0.052 | sN | 0.1146 | 0.088 | 91.8 | 0.116 | 95.1 |
| 2.1 | 7 | 0.052 | N | 0.0692 | 0.066 | 94.1 | 0.070 | 95.2 |
| 2.2 | 7 | 0.052 | $t_5$ | 0.0694 | 0.067 | 94.3 | 0.070 | 95.1 |
| 2.3 | 7 | 0.052 | U | 0.0689 | 0.066 | 94.2 | 0.068 | 94.8 |
| 2.4 | 7 | 0.052 | sN | 0.0693 | 0.066 | 94.0 | 0.070 | 95.1 |
| 3.1 | 3 | 0.026 | N | 0.0613 | 0.043 | 89.0 | 0.063 | 95.3 |
| 3.2 | 3 | 0.026 | $t_5$ | 0.0626 | 0.044 | 89.4 | 0.063 | 95.1 |
| 3.3 | 3 | 0.026 | U | 0.0607 | 0.043 | 89.4 | 0.061 | 95.1 |
| 3.4 | 3 | 0.026 | sN | 0.0617 | 0.043 | 88.8 | 0.063 | 95.2 |
| 4.1 | 3 | 0.013 | N | 0.0393 | 0.032 | 90.3 | 0.040 | 95.2 |
| 4.2 | 3 | 0.013 | $t_5$ | 0.0408 | 0.038 | 94.0 | 0.040 | 94.8 |
| 4.3 | 3 | 0.013 | U | 0.0367 | 0.032 | 92.1 | 0.037 | 95.0 |
| 4.4 | 3 | 0.013 | sN | 0.0395 | 0.032 | 90.2 | 0.041 | 94.9 |

Considering first the computations of the median, $m(Y)$, in cases 1, 2 and 3 of the correction term, the normal, t and uniform distributions are symmetric, as is the distribution of $X$ in all cases, so medians are equal to means. And because the measurement model is linear means are propagated exactly in the GUF, CUF and Monte Carlo. All methods correctly give $m(Y) = 5.7120$. The exception is the skew-normal distribution for $C$ in case 4, which has mean zero but median $m(C) = -0.00462$. For each of cases 1.4, 2.4, 3.4 and 4.4 the GUF computes the mean of $Y$ to be 5.7120, and this is inferred also to be the median. In those same cases, the CUF computes the median to be $m(Y) = 5.7120 - 0.0046 = 5.7074$. The exact median of $Y$, computed by Monte Carlo, is 5.7109 in cases 1.4 and 2.4, 5.7098 in case 3.4 and 5.7087 in case 4.4.

When the model includes an input with an asymmetric distribution, neither GUF nor CUF computes the median of $Y$ exactly. Both are approximate, and we see that GUF is more accurate when the skewed input $C$ has lower uncertainty than that for the symmetric input $X$, while CUF is more accurate when the skewed input has higher uncertainty. However, in all cases the errors in computing $m(Y)$ are very small compared with the uncertainty in $Y$. This example supports the assertions in Subsections 4.1 and 4.3 that both GUF and CUF are applicable if 'the input distributions making substantial contributions to the uncertainty in $Y$ have a symmetric (or almost symmetric) form similar to a normal or t distribution'.

The performance of the GUM uncertainty framework (GUF) and of our proposed characteristic uncertainty framework (CUF) in computing the characteristic uncertainty $c(Y)$ of the measurand are shown in the last five columns of table 2. For each case we show in columns 5, 6 and 8 respectively the 'true' $c(Y)$ values from MCM and the $c(Y)$ values given by the GUF and CUF. Columns 7 and 9 show the percentage coverage, computed using MCM, for the implied 95 % intervals $m(Y) \pm 2c(Y)$ from GUF and CUF.

Considering first the figures for the CUF in the last two columns of table 2 we note the following:

- Propagation of characteristic uncertainties using CUF produces in every case a $c(Y)$ very close to the true value from MCM. In this example, therefore, the transferability of characteristic uncertainties is affirmed.
- Furthermore, the true coverage of the CUF's implied 95 % intervals is seen in every case to be very close to 95 %.
- The various cases for $C$ (normal, t, uniform or skew-normal) make little difference to the accuracy of the approximations. They have the biggest influence in the last block of the table, Cases 4.1 to 4.4, when $u(X) = 0.013$ and there is therefore more uncertainty about $C$ than $X$, which can happen occasionally in practice.

This example therefore supports our claim that the CUF provides a good approximation to the true median and characteristic function in a case where the conditions for its applicability hold.

In columns 6 and 7 of table 2 propagation according to the GUM, using the Welch-Satterthwaite approximation is seen to be less accurate. The GUM values for $c(Y)$ are invariably smaller than the true values, with coverage appreciably less than 95 %. Similar findings of inadequate coverage of intervals based on the Welch-Satterthwaite approximation have been reported elsewhere [20], but it should be noted that these findings are from a Bayesian perspective, under which the MCM provides exact computation of the Bayesian posterior distribution of $Y$. From the frequentist perspective, coverage of the GUF 95 % interval should be judged instead on the basis of very many repetitions of the measurement, and Welch-Satterthwaite has been shown to be a good approximation with coverage typically close to 95 % when its assumptions hold [10]. However, those assumptions do not generally hold when Type B evaluations are involved.

Our position is that only the Bayesian paradigm properly allows the combination of Type A and Type B evaluations, and that the MCM computation is indeed the gold standard against which other methods should be judged.

EXAMPLE 2 *Six-term model*

The Standards Publication CEN/TR 16988:2016 [1] is entitled 'Estimation of uncertainty in the single burning item test'. Clause 2.5.13.2 deals with the uncertainty concerning an input described as the 'velocity profile correction factor', which we will denote by $\kappa$ and which is expressed using the submodel

$$\kappa = \frac{1}{5}\sum_{i=1}^{5}\frac{v_i}{v_c} \tag{1}$$

with six input quantities. $v_i$, $i = 1,\ldots,5$, are velocity measurements taken on five different radii and $v_c$ is a central measurement. Each measurement is actually the average of four indications taken at 90° intervals. These measurements are reported in table 3. The characteristic uncertainty of each input is the standard uncertainty multiplied by $k_3/2 = 1.591$.

**Table 3:** Input data, Example 2

| Quantity | Estimate /ms$^{-1}$ | Standard uncertainty/ms$^{-1}$ | Degrees of freedom | Characteristic uncertainty/ms$^{-1}$ |
|---|---|---|---|---|
| $v_1$ | 7.00 | 1.132 | 3 | 1.801 |
| $v_2$ | 9.39 | 0.412 | 3 | 0.656 |
| $v_3$ | 10.62 | 0.531 | 3 | 0.845 |
| $v_4$ | 11.25 | 0.180 | 3 | 0.286 |
| $v_5$ | 12.37 | 0.233 | 3 | 0.355 |
| $v_c$ | 12.39 | 0.636 | 3 | 1.012 |

We will denote an estimate by placing a hat over the quantity, so that for instance $\widehat{v}_1 = 7.00\,\mathrm{ms}^{-1}$. Following the GUF, the estimate of $\kappa$ is obtained by substituting the estimates of the input quantities into the measurement function, giving $\widehat{\kappa} = 0.817$. However, to obtain the standard uncertainty and

expanded uncertainty, the model (1) is linearized by expanding in a first order Taylor series around the estimated values of the quantities, which yields

$$\begin{aligned}\kappa &= \widehat{\kappa} + \frac{1}{5\widehat{v}_\mathrm{c}}\sum_{i=1}^{5}(v_i - \widehat{v}_i) - \frac{1}{5\widehat{v}_\mathrm{c}^2}(v_\mathrm{c} - \widehat{v}_\mathrm{c})\sum_{i=1}^{5}\widehat{v}_i \\ &= 0.817 + 0.016\,142\sum_{i=1}^{5}(v_i - \widehat{v}_i) - 0.065\,940(v_\mathrm{c} - 12.39)\,.\end{aligned} \tag{2}$$

For this example, we will simply use the linearized version (2) as the measurement model, but we will return to the original nonlinear model (1) in Appendix A. Because all the inputs are symmetric, their estimates are also means and medians. Both the GUF and CUF will correctly infer the true $m(Y) = 0.817$.

The GUF now applies the LPU to the standard uncertainties of the inputs to obtain the standard uncertainty $u(\kappa) = 0.0473$. Next, the Welch-Satterthwaite formula gives 4.66 degrees of freedom for $\kappa$. Therefore the characteristic uncertainty is obtained as $c(\kappa) = 0.0473\,k_{4.66}/2 = 0.0622$. The CUF instead applies the LPU to the characteristic uncertainties, resulting in $c(\kappa) = 0.0752$.

For comparison, we apply the Monte Carlo method to (2). We obtain $c(\kappa) = 0.0761$. The true coverage probabilities for the implied 95 % intervals are 91.6 % for the GUM uncertainty framework and 94.8 % for the characteristic uncertainty framework. This example therefore lends further support to the indication from Example 1, that simple propagation of characteristic uncertainties of the model inputs yields an accurate approximation to the true characteristic uncertainty of the measurand, with close to 95 % coverage, and that from the Bayesian perspective the GUM uncertainty framework is less accurate.

EXAMPLE 3 *Sum of skewed inputs*

Our third example illustrates how in some extreme situations the CUF may perform less well, due to the way propagation of medians for skewed distributions may misrepresent the median of the measurand.

Consider the model

$$Y = \sum_{i=1}^{M} X_i\,,$$

where the measurand is the sum of $M$ inputs. Suppose for convenience in this example that the $X_i$ all have Type A evaluations based on samples of $n = 6$ normal observations, and all have sample means $\bar{x} = 1$ and frequentist standard uncertainties $u_\mathrm{f}(\bar{x}) = 0.8$.

The standard GUM analysis in this case is straightforward. The estimate of $Y$ is $y = M$, with standard uncertainty $u_\mathrm{f}(y) = 0.8\sqrt{M}$. The Central Limit Theorem says that for large $M$ the sum of independent random variables has a normal distribution asymptotically, and because the t distributions are unimodal and symmetric this will apply even for moderate $M$. This is supported by application of the Welch-Satterthwaite formula, which gives an effective degrees of freedom of $d = 5M$, and therefore for any $M \geq 4$ the implied characteristic uncertainty is $c(y) = 0.98 \times 0.8\sqrt{M} = 0.784\sqrt{M}$.

We now introduce a condition that it is known that all the $X_i$ are necessarily positive. Individually, an estimate of 1 with standard uncertainty 0.8 and 5 degrees of freedom would lead to an expanded uncertainty of 2.0565 and an implied 95 % coverage interval from −1.0565 to 3.0565, which includes negative values in contradiction of the constraint. Although there may be alternative frequentist analyses to take account of this constraint, it would not be deemed a problem in practice since for even moderate $M$ the standard uncertainty $u(y)$ will be small enough for no such issues to arise. For instance, with $M = 4$ the 95 % coverage interval $4 \pm 3.136$ is entirely positive.

Now applying a Bayesian Type A analysis the constraint is simple to apply. The posterior t distribution is truncated to positive values of $X_i$. The truncated t distribution is shown in figure 5. This distribution has mean $E(X_i) = 1.2543$ and standard uncertainty $u_\mathrm{b}(X_i) = 0.8143$. However, its median is $m(X_i) = 1.1413$ and its characteristic uncertainty is $c(X_i) = 0.7803$.

The exact Bayesian measurement result for $Y$ can be computed by MCM, with mean $E(Y) = 1.2543M$ and standard uncertainty $u(Y) = 0.8143\sqrt{M}$. Again for $M \geq 4$ the distribution will be very close to

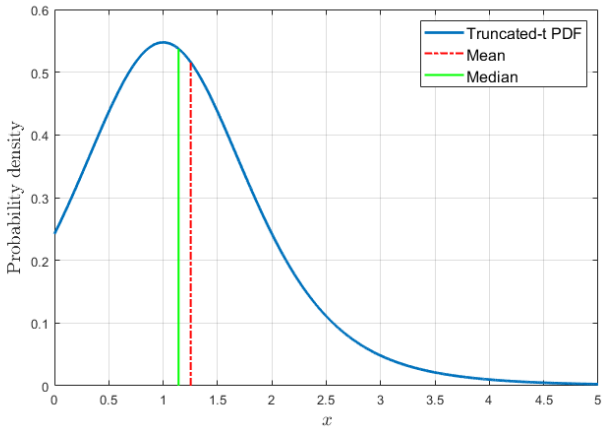


**Figure 5:** Density function of truncated-t distribution

a normal distribution, so the median is the same as the mean, $m(Y) = 1.2543M$ and the characteristic uncertainty is $c(Y) = 0.98 \times 0.8143\sqrt{M} = 0.7980\sqrt{M}$. However, applying the CUF the median is estimated as $1.1413M$ and the characteristic uncertainty is estimated as $0.7803\sqrt{M}$ For sufficiently large $M$ the CUF estimates will deviate substantially from the exact Bayesian values.

Table 4 presents some calculations for $M = 4, 9$ and 16. There is little difference between the three characteristic uncertainty values for any given $M$, but the various median values deviate systematically from each other, and these differences become relatively larger compared with the characteristic uncertainty as $M$ increases. This is shown in the percentage coverages for GUF and CUF in columns 6 and 9. These are calculated using the corresponding 95 % intervals $m(Y) \pm 2c(Y)$ and the gold standard normal distribution from MCM.

**Table 4:** Comparing GUM and characteristic uncertainty frameworks, Example 3

| $M$ | MCM $m(Y)$ | MCM $c(Y)$ | GUF $m(y)$ | GUF $c(Y)$ | GUF % | CUF $m(Y)$ | CUF $c(Y)$ | CUF % |
|---|---|---|---|---|---|---|---|---|
| 4 | 5.017 | 1.596 | 4 | 1.568 | 89.8 | 4.565 | 1.561 | 93.5 |
| 9 | 11.289 | 2.394 | 9 | 2.352 | 83.6 | 10.272 | 2.341 | 92.3 |
| 16 | 20.069 | 3.192 | 16 | 3.136 | 75.1 | 18.261 | 3.121 | 90.7 |

Looking first at the CUF percentages in column 9, we see that coverage steadily decreases from the nominal 95 % as $M$ increases. At $M = 16$ it is 90.7 %, which may be regarded as unacceptably low. The explanation is that in this example the conditions we have identified for the CUF to be applicable do not hold. The distribution of each $X_i$, shown in figure 5, is markedly skew. The combination of many such skew distributions, all skewed in the same direction, causes the accumulating error in the estimated median. For small $M$, the error is small compared with the uncertainty in $Y$, and the CUF median and characteristic uncertainty remain useful and meaningful expressions for the recipient of the measurement result.

Although this example illustrates the failure of the CUF to give acceptable approximations to the true Bayesian median and coverage interval, it is comforting that it only arises when a relatively large number of inputs, all appreciably skewed, are combined. We believe that practical instances of such a measurement model will be rare.

Turning to the GUF percentages in column 6 of table 4, they suggest that the GUF coverage is unacceptably low even for $M = 4$. Nevertheless, this is not the case from a frequentist perspective. The estimate $y = M$ is unbiased, its sampling standard deviation is validly estimated as $0.8\sqrt{M}$ and $M \pm 0.784\sqrt{M}$ is an exact 95 % confidence interval. If very many repetitions of the measurement were

performed and the interval computed each time then 95 % of those intervals would contain the true value of the measurand. From the frequentist perspective, the MCM is not a gold standard; it computes the Bayesian measurement result exactly, but this differs from what is a valid frequentist result.

The difference between the frequentist and Bayesian analyses arises from the fact that the Bayesian posterior distribution for $X_i$ implements the known constraint that $X_i \geq 0$, and this leads to a posterior expectation that $X_i$ is more likely to be above the sample mean $\bar{x} = 1$ than below 1. The reasoning behind this expectation is as follows. Consider that $\bar{x} = 1$ could have arisen from a true value $X_i$ greater than 1 and a negative average measurement error, or $X_i$ less than 1 and a positive average error. A positive error of a given size has the same probability as a negative error of that size. Therefore given $\bar{x} = 1$ it is equally likely for $X_i$ to be 1.5 or 0.5, for example, but is not equally likely to be 2.5 or $-0.5$, since the latter is ruled out by the constraint. It is here that the asymmetry in the posterior distribution is created, leading to a larger probability for each $X_i$ to be above 1 than below 1. This effect would apply for any value of $\bar{x}$, but is non-trivial in this instance because the sampling error is relatively large compared with $\bar{x}$.

We remain convinced that the Bayesian paradigm is the more appropriate methodology for metrology.

In the first two examples, the conditions outlined in Subsections 4.1 and 4.3 for applicability of the GUF and CUF are satisfied, namely that the models are linear or almost linear, and that the probability distributions are close to the normal or t forms and nearly symmetric for all inputs making a substantial contribution to the uncertainty in the measurand. Full conditions for the valid applicability of the GUF are given in [7, clauses 5.7 and 5.8].

The examples confirm that under these conditions the characteristic uncertainty framework provides accurate evaluation of the median and characteristic uncertainty of a measurand, and that from the Bayesian perspective it is more accurate than the GUM uncertainty framework. More testing would certainly be warranted to add further confirmation.

The third example concerns a rare situation where the conditions for the applicability of the CUF do not hold, involving a measurement model with many markedly skew input distributions. In such a situation, the error in the CUF propagation of the median may be sufficient for the implied 95 % interval to have poor coverage despite the characteristic uncertainty being propagated accurately.

Methods of propagation similar to the CUF have been suggested by other authors. Williams [23] and Kacker [13], noting how closely the Bayesian standard uncertainty approximates the characteristic uncertainty in the case of a normal sample (see Subsection 3.2), propose simply propagating the Bayesian standard uncertainty using the LPU and then assuming the distribution of $Y$ is normal. Their suggestion approximates to the CUF in this case, but leads to less accurate propagation and is less generally applicable.

CUF's propagation of characteristic uncertainties is equivalent to propagating expanded uncertainties. The GUM [3, clause E.3.3] points out that it is legitimate to propagate fixed multiples of standard uncertainties using the LPU, but this would not apply to propagating variable multiples, such as expanded uncertainties. Nevertheless, in the original analysis of the six-term model (1) [1], expanded uncertainties are propagated through the linearized model (2) in this way without comment.

### 4.5 Propagation guidelines

The Monte Carlo method is the gold standard for propagating input uncertainty through all kinds of measurement models to compute uncertainty about a measurand. Nevertheless, the GUM Uncertainty Framework is still by far the more widely used method in laboratory practice. MCM is more complex to apply, requiring some computing power and expertise. And

although the GUF is only recommended for models that are linear or close to linear, the linearization technique is very attractive, and so it is often used even in markedly nonlinear models.

The comparison between the GUF and CUF approaches in the two examples suggest the following conclusions.

- The characteristic uncertainty framework is simpler to apply than the GUM uncertainty framework, because it does not entail the computation of a degrees of freedom through the Welch-Satterthwaite formula.
- In linear or nearly linear models, the CUF's simple propagation of characteristic uncertainties using the LPU generally produces an accurate approximation to the true characteristic uncertainty of the measurand, as computed by MCM, with true coverage close to 95 %.
- In linear or nearly linear models, the GUF appears to yield less accurate approximation of the true characteristic uncertainty, with coverage that is typically lower than the claimed 95 %.

We argue, therefore, that wherever the GUF is applicable the characteristic uncertainty framework should be seriously considered as a more viable method of propagation. There remains no compelling reason to retain the use of standard uncertainty in metrology.

We have proposed in Subsection 3.5 that the probability distribution of the measurand should always be reported as the primary measurement result. When the CUF has been used to compute the median and characteristic uncertainty of $Y$, and therefore the appropriate conditions apply, it will be adequate to report a normal distribution.

When the CUF is not applicable, for instance when the model is markedly nonlinear or when there are inputs with markedly asymmetric distributions that make a substantial contribution to the uncertainty in the measurand, we would always recommend the Monte Carlo method if the appropriate tools and expertise are available.

# 5 Conclusions

When reporting a measurement result for a quantity, it is important to express the metrologist's knowledge fully in the form of a probability distribution. However, it is equally important to provide useful and meaningful summaries of that information for the benefit of the recipient of that result. The median and characteristic uncertainty should be the primary summaries included in the measurement result. A plot of the PDF of the distribution is also valuable as a visual summary, while other summaries may also be useful depending on context, or where the distribution is markedly skew (as discussed in Appendix C).

We find no value in reporting the standard uncertainty (standard deviation), because it is not a meaningful summary, may not exist and may give a misleading impression in the case of a distribution with heavy tails (low degrees of freedom). Furthermore, conflicting definitions of the standard uncertainty have given rise to confusion and friction. Characteristic uncertainty may defuse that debate.

When a quantity of interest (measurand) is expressed through a measurement model in terms of one or more input quantities, a procedure is needed for computing the distribution and summaries for the measurand in terms of the corresponding properties of the inputs. The gold

standard for this propagation from the Bayesian perspective is the Monte Carlo method as proposed in the GUM Supplement 1. Given the (joint) distribution of the inputs, it yields the distribution of the measurand in the form of a large sample from that distribution. The distribution may be reported in this form, as an electronic file, or as a suitable standard statistical distribution that is a good approximation fitted to the sample. Summaries such as median and characteristic uncertainty may be computed directly from the sample. The PDF plot may be a kernel density plot derived from the sample or a plot of a fitted distribution.

Provided that the model is linear or nearly linear, and that all inputs making substantial contributions to the uncertainty in the measurand have symmetric or nearly symmetric distributions similar to a normal or t distribution, the characteristic uncertainty framework (CUF) may be used to compute good approximations to the median and characteristic uncertainty of the measurand. In that case, the distribution of the measurand may be reported as the normal distribution matching those summaries.

The GUM uncertainty framework, as set out in the GUM and its Supplement 1, relies on analogous conditions to the CUF for its validity, and appears to be no more accurate when compared with the precise computations from the Monte Carlo method. Indeed, in all the examples we have explored its coverage, computed from the Bayesian perspective, seems to be consistently below the nominal 95 %. We therefore see no useful role for the standard uncertainty in propagation that is not equally served by the characteristic uncertainty. Moreover, on the basis of a number of examples, the coverage provided by the CUF is very close to 95 %, whereas from the Bayesian perspective that produced by GUF can be appreciably less.

Our principal, and most radical, recommendation is that the characteristic uncertainty should be the primary single-figure expression of uncertainty in measurement.

## Acknowledgements

This work was supported by an ISCF (Industrial Strategy Challenge Fund) Metrology Fellowship grant provided by the UK government's Department for Business, Energy and Industrial Strategy (BEIS).

The authors are also grateful for helpful comments from participants at meetings where these ideas have been aired.

## A Infinite standard deviations

We discuss here situations in which the standard uncertainty of a quantity may be infinite, including instances where the mean does not exist. These cases will cause insurmountable problems if measurement uncertainty is defined to be a standard uncertainty, and if the estimate of a quantity is required to be the mean. We emphasise that no such problems arise with the median and characteristic uncertainty. These summaries exist in all such cases, in addition to being well-defined, clear and meaningful for the recipient of a measurement result.

Probability distributions with infinite standard deviations arise in a number of ways, one example being GUM-S1's Bayesian Type A evaluation for a normal sample discussed in Subsection 4.4.

As stated in Subsection 2.2, a Bayesian Type A evaluation combines information in the data with prior information, and it is the standard deviation of the posterior distribution that is

the Bayesian standard uncertainty. In this example GUM-S1 uses a 'noninformative' prior distribution that is supposed to represent a null state of prior knowledge. This is a common and frequently useful device in Bayesian analyses generally, but when the information in the data is very limited a 'noninformative' prior distribution can lead to a posterior distribution with infinite standard deviation. This is the situation with the GUM-S1 analysis of the normal sample with $n < 4$. Indeed, when $n = 2$ neither the standard deviation nor the expectation of the measurand exists.

Although the median and characteristic uncertainty resolve such problems, it also worth noting that a situation of 'no prior information' is unrealistic. Before carrying out a measurement, the metrologist will have some prior expectations regarding the quantity to be measured and the error characteristics of the measuring system. One reason for the use of a 'noninformative' prior distribution by GUM-S1 is that the use of the metrologist's subjective prior information is controversial and may in some contexts be unacceptable. In the case of a sample of $n < 4$ from a normal distribution, even a small amount of prior information suffices to produce a posterior distribution with a finite Bayesian standard uncertainty. Cox and O'Hagan (paper in development) show that relatively weak prior information about the measurement variance $\sigma^2$, such as might normally be expected to be available quite uncontroversially, will yield not only a finite posterior standard deviation but also a material reduction in the length of a coverage interval (also see [8]).

Infinite standard uncertainties can also arise due to the nature of the measurement model. This may occur when a measurand is expressed as a ratio of two inputs. Wesson, Stock and Scicluna [22] discuss the flux ratio of doubly ionised oxygen emission lines, arising at wavelengths of 500.7 nm and 495.9 nm:

$$V = F_{500.7}/F_{495.9}\,.$$

If the denominator has a Type A evaluation resulting in it having a normal or t distribution, then the distribution of $V$ has neither a mean nor a standard deviation, due to the possibility of $F_{495.9}$ being arbitrarily close to zero. In practice, the uncertainty in $F_{495.9}$ may be small, such that the probability of being in the neighbourhood of zero is tiny, but the distribution of $V$ will nevertheless have infinite standard uncertainty.

The same situation arises when the measurand $X$ is modelled as a ratio of differences. For example, a coefficient of expansion $X$ may be measured by the ratio of change in length to change in temperature

$$X = \frac{L_1 - L_0}{T_1 - T_0}\,.$$

Given a sample of indications of $T_1$ and $T_0$, even though the sample is large and the relative uncertainty around the difference $T_1 - T_0$ is small, there is still in principle a nonzero probability that it might be negative. The result is that the standard uncertainty of $X$ is infinite and its mean is undefined.

In situations such as these, application of the GUF will not reveal the fact that the mean of the measurand is undefined or that the standard uncertainty is infinite. It will yield an estimate that is supposed to approximate to the mean and a finite combined standard uncertainty.

A Bayesian analysis with noninformative prior distributions will have the same effect. Furthermore, application of the Monte Carlo method as advocated in GUM-S1 will always erro-

neously yield a value for the mean and a finite standard uncertainty for $X$ based on a finite Monte Carlo sample.

In both cases, the problem arises from representing quantities that are necessarily positive by uncertainty distributions that fail to respect the constraint. For instance, it does not occur when working with logarithms, so that the implied distributions are lognormal or log-t. Nor does it occur when the prior knowledge of the constraint is properly represented in an informative prior distribution. However, the routine application of the GUM's Type A evaluation for a normal sample for quantities that are necessarily positive is widespread in metrology, with the result that computations of standard uncertainties may be highly unreliable.

Consider Example 2 in Subsection 4.4. The division by $v_c$ in the model (1) also results in $\kappa$ having infinite variance and an undefined mean. The median and characteristic uncertainty are nevertheless well defined and can for instance be computed to any desired accuracy by the Monte Carlo method. A Monte Carlo sample of size $10^6$ reported a mean for $\kappa$ of 1.3 and a standard deviation of 482.5. These numbers are completely spurious, would change substantially if we took another $10^6$ samples, and would never converge no matter how large a sample we generated. (Indeed, the nature of the problem would be identified by the non-convergence of the adaptive method in GUM-S1 for evaluating the mean and standard uncertainty to a target numerical accuracy.) The same $10^6$ sample reported a median of $m(\kappa) = 0.8173$ and the characteristic uncertainty was found to be $c(\kappa) = 0.076$. These are close to the MCM and CUF figures given in Subsection 4.4, confirming the applicability of the CUF even in a model of this degree of nonlinearity. In contrast, the nonlinearity in this model is catastrophic if we insist on using standard uncertainty.

# B COMPUTING THE MEDIAN AND CHARACTERISTIC UNCERTAINTY

We present various ways to derive the median and characteristic uncertainty of a quantity.

## B.1 COMPUTATION BY MONTE CARLO

A generic technique, that can be used in several of the contexts described below, is computation by a Monte Carlo method. Suppose that we have a sample of $M$ values of $X$, drawn randomly from its probability distribution. First, arrange them in non-decreasing order, $x_{[1]} \le \ldots \le x_{[M]}$. If $M$ is an odd number, the median is $m(X) = x_{[(M+1)/2]}$, otherwise $m(X) = (x_{[M/2]} + x_{[(M/2)+1]})/2$.

Now define $y_{[1]} \le \cdots \le y_{[M]}$ computed by taking all the values $|x_{[i]} - m(X)|$ for $i = 1, \ldots, M$ and arranging them in non-decreasing order. Then $c(X) = y_{[t]}/2$, where $t = 0.95(M+1)$, rounded up if necessary to the next integer.

For these computations to be sufficiently accurate in practice we recommend $N \ge 10^6$; see also [7, clause 7.2].

## B.2 COMPUTATIONS FOR A SINGLE EVALUATION

Suppose now that the distribution for $X$ arises from a single Type A or Type B evaluation.

We have of course considered in some detail the most widely used Type A evaluation, namely that involving a sample of size $n$ from a normal distribution. In this case the distribution of $X$ is a t distribution, the median is the sample mean, $m(X) = \bar{x}$, and $c(X) = k_{n-1}s/(2\sqrt{n})$.

If the distribution of $X$, whether obtained as the posterior distribution from a Bayesian Type A evaluation or as the metrologist's judgement in a Type B evaluation, has the form of a standard

probability distribution in statistics, then there may be explicit expressions for the median and/or characteristic uncertainty (possibly involving functions whose values can be looked up from tables or computed by standard software). The t distribution is one example. Another is the half-normal distribution featured in Appendix C; the HN$(a, b^2)$ distribution has $m(X) = a + 0.6745b$ and $c(X) = 0.6427b$.

In cases where the distribution has a standard form but does not have explicit formulae for median and/or characteristic uncertainty, there are two simple computational techniques to evaluate them. One is to use numerical integration to compute the cumulative distribution function $G(x)$ at any $x$, and then numerical methods to solve $G(m(X)) = 0.5$ and $G(m(X) + 2c(X)) - G(m(X) - 2c(X)) = 0.95$.

The second technique is to draw a random sample of $M$ values from the distribution and then apply the Monte Carlo computation given in Appendix B.1.

Finally, some Bayesian Type A evaluations will not give a posterior distribution of a standard form such that numerical integration or direct sampling is possible. Instead, Markov chain Monte Carlo [21] is another, rather more complex, tool to obtain a sample of $M$ values from the distribution.

### B.3 Computations for a measurement model

Now suppose that the measurand $X$ is expressed through a measurement model in terms of a number of input quantities. As described in Section 4.5, the Characteristic Uncertainty Framework provides a simple way to compute the median and characteristic uncertainty of $X$ approximately from those of the inputs, while the Monte Carlo method of Section 4.2 does so to any desired accuracy.

## C Skew distributions

Three example distributions were presented in Section 3.4. All would fit the criterion of Sections 4.1 and 4.3 for the GUF and CUF to be applicable, namely that they have 'a symmetric (or almost symmetric) form similar to a normal or t distribution'. The family of skew-normal distributions includes distributions that are far from symmetric (as their name suggests), but the case presented in Section 3.4 has only moderate skewness. The same is true of the gamma family of distributions; gamma distributions can be markedly skew but the example in Section 3.4 is only modestly so.

To illustrate the use of median and characteristic uncertainty in more strongly asymmetric distributions, we present two more examples.

**Half-normal distribution.** The half-normal distribution is the same as the normal distribution except that the density is reduced to zero for all values below what would have been the mean of the normal distribution. It is a particular case of a truncated normal distribution, which can be appropriate as a judgement distribution when the value of the quantity is bounded either above or below (or both). We now suppose that a metrologist judges, based on the available evidence, that $X$ is necessarily non-negative, and represents uncertainty about $X$ with the half-normal distribution HN$(0, 0.0481^2)$. The density is shown in figure 6. It has mean $\mathrm{E}(X) = 0.0384$, standard deviation $u(X) = 0.0290$, median $m(X) = 0.0324$ and characteristic uncertainty $c(X) = 0.0309$. The skewness is more marked for this distribution and there are larger differences between $m(X)$ and $\mathrm{E}(X)$, and between $c(X)$ and $u(X)$ than we saw in the examples of Section 3.4.

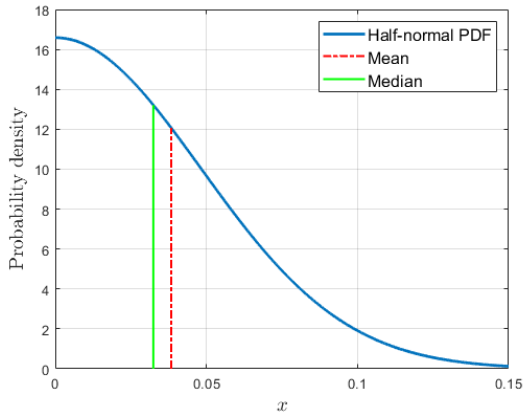


**Figure 6:** Density function of half-normal distribution

**Lognormal distribution.** A lognormal distribution can be another representation of a metrologist's uncertainty about a quantity that must be non-negative. It can also arise in Type A evaluation when the errors in the sample of indications are believed to follow a normal distribution on the log scale. Suppose that $X$ has the lognormal distribution $\mathrm{lN}(-4.311, 1)$ so that $\ln X$ has the normal distribution with mean $-4.311$ and variance 1. As shown in figure 7, this distribution is strongly skewed. It has mean $\mathrm{E}(X) = 0.0221$, standard deviation $u(X) = 0.0290$, median $m(X) = 0.0134$ and characteristic uncertainty $c(X) = 0.0281$. The mean is no longer a useful estimate because $X$ is twice as likely to be below 0.0221 as to be above it. The median value of 0.0134 can be seen as a more representative value for $X$.

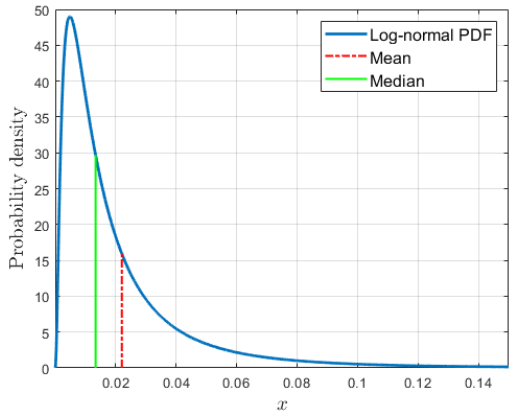


**Figure 7:** Density function of lognormal distribution

As the distribution of $X$ becomes increasingly skew, the mean is found further into the long tail of the distribution and is increasingly unrepresentative as an estimate of $X$. The median value is always a more representative estimate and its status as a central value, with equal probability for $X$ to be above or below the median, gives it a clear and unambiguous interpretation.

However, when the distribution has appreciable skewness some care may be required when interpreting $c(X)$ as a measure of uncertainty. For instance, in both these examples the interval $m(X) \pm 2c(X)$ extends below zero, which is not ideal for a quantity $X$ that cannot be negative. In the case of the lognormal example, for instance, $m(X) - 2c(X) = -0.0428$. It remains true that

there is a 95 % probability that $X$ will lie in the range $m(X) \pm 2c(X) = (-0.0428, 0.0696)$. The meaning of characteristic uncertainty is not affected, but in this case there is clearly a 95 % probability that $X$ will lie in the narrower interval of $(0, 0.0696)$.

We suggest that in such cases the recipient would benefit from being given additional summaries of the distribution.

It will generally be useful to present a plot of the PDF of the distribution, In the examples above, figures 6 and 7 show the skewness clearly and will aid the recipient's interpretation of the median and characteristic uncertainty. Indeed, we recommend that a PDF plot should form a standard component of the measurement result. Even when the distribution has 'a symmetric (or almost symmetric) form similar to a normal or t distribution', a PDF plot such as in figures 2. 3 and 4 provides a meaningful visual summary of the distribution, showing which values of $X$ are more or less probable.

Various quantitative summaries of skewness can also be proposed, but may be of limited practical value. Although skew distributions arise occasionally in practice – Possolo et al. [19] give some examples – they are not covered explicitly in the GUM and have received little attention in the metrology literature.

## REFERENCES

[1] PD CEN/TR 16988:2016, *Estimation of uncertainty in the single burning item test*.

[2] A. Azzalini. A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985.

[3] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology, JCGM 100:2008.

[4] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Supplement 2 to the "Guide to the expression of uncertainty in measurement" — Models with any number of output quantities. Joint Committee for Guides in Metrology, JCGM 102:2011.

[5] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Guide to the expression of uncertainty in measurement — Part 6: Developing and using measurement models. Joint Committee for Guides in Metrology, GUM-6:2020.

[6] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. International Vocabulary of Metrology — Basic and General Concepts and Associated Terms. Joint Committee for Guides in Metrology, JCGM 200:2012.

[7] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method. Joint Committee for Guides in Metrology, JCGM 101:2008, 2008.

[8] M Cox and K Shirono. Informative Bayesian Type A uncertainty evaluation, especially applicable to a small number of observations. *Metrologia*, 54(5):642–652, 2017.

[9] William F. Guthrie. NIST/SEMATECH e-Handbook of Statistical Methods (NIST Handbook 151), 2020.

[10] B D Hall and R Willink. Does "Welch-Satterthwaite" make a good uncertainty estimate? *Metrologia*, 38(1):9–15, 2001.

[11] P M Harris, C E Matthews, M G Cox, and A B Forbes. Summarizing the output of a Monte Carlo method for uncertainty evaluation. *Metrologia*, 51(3):243, 2014.

[12] R. Kacker and A. Jones. On use of Bayesian statistics to make the Guide to the Expression of Uncertainty in Measurement consistent. *Metrologia*, 40:235–248, 2003.

[13] Raghu N. Kacker. Bayesian alternative to the ISO-GUM's use of the Welch-Satterthwaite formula. *Metrologia*, 43(1):1–11, 2005.

[14] Ignacio Lira and Wolfgang Wöger. Comparison between the conventional and Bayesian approaches to evaluate measurement data. *Metrologia*, 43:S249–S259, 2006.

[15] R.G. Miller. Beyond ANOVA, Basics of Applied Statistics. *Biometrical Journal*, 30(7):874, 1988.

[16] A. O'Hagan and T. Leonard. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63(1):201–203, 1976.

[17] Anthony O'Hagan. Eliciting and using expert knowledge in metrology. *Metrologia*, 51(4):S237, 2014.

[18] Antonio Possolo and Hari K. Iyer. Concepts and tools for the evaluation of measurement uncertainty. *Review of Scientific Instruments*, 88(1):011301, 2017.

[19] Antonio Possolo, Christos Merkatas, and Olha Bodnar. Asymmetrical uncertainties. *Metrologia*, 56(4):045009, 2019.

[20] D Turzeniecka. Comments on the accuracy of some approximate methods of evaluation of expanded uncertainty. *Metrologia*, 36(2):113–116, 1999.

[21] Don van Ravenzwaaij, Pete Cassey, and Scott D. Brown. A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin & Review*, 25(1):143–154, 2016.

[22] R. Wesson, D. J. Stock, and P. Scicluna. The probability distribution functions of emission line flux measurements and their ratios. *Monthly Notices of the Royal Astronomical Society*, 459(4):3475–3481, 2016.

[23] Alex Williams. An alternative to the effective number of degrees of freedom. *Accreditation and Quality Assurance*, 4(1-2):14–17, 1999.