Software Support for Metrology
Best Practice Guide No. 6

# Uncertainty and Statistical Modelling

*M G Cox, M P Dainton and P M Harris*

March 2001

# Software Support for Metrology
# Best Practice Guide No. 6

# Uncertainty and Statistical Modelling

M G Cox, M P Dainton and P M Harris
Centre for Mathematics and Scientific Computing

March 2001

ISSN 1471–4124



Authorised by Dr Dave Rayner,
Head of the Centre for Mathematics and Scientific Computing

National Physical Laboratory,
Queens Road, Teddington, Middlesex, United Kingdom TW11 0LW

# ABSTRACT

This guide provides best practice on the evaluation of uncertainties within metrology, and on the support to this topic given by statistical modelling. It is motivated by two principle considerations. One is that although the primary guide on uncertainty evaluation, the Guide to the Expression of Uncertainty in Measurement (GUM), published by ISO, can be expected to be very widely applicable, the approach it predominantly endorses contains some limitations. The other is that on the basis of the authors' considerable contact with practitioners in the metrology community it is evident that important classes of problem are encountered that are subject to these limitations. These problems include

- Measurements at the limits of detection, where the magnitude of an uncertainty is comparable to that of the measurand
- Measurements of concentrations and other quantities that are to satisfy conditions such as summing to 100%
- Obtaining interpolated values and other derived quantities from a calibration curve, accounting for the correlations that arise
- Determining "flatness", "roundness" and other such measures in dimensional metrology, whilst avoiding the anomalous "uncertainty behaviour" that is sometimes observed
- Treating the asymmetric distributions that arise when dealing with the magnitudes of complex variables in acoustical, electrical and optical work.

There would appear to be inadequate material available in existing guides to facilitate the valid solution to such problems.

Central to consideration is the need to carry out uncertainty evaluation in as scientific a manner as economically possible. Although several approaches to uncertainty evaluation exist, the GUM has been very widely adopted (and is strongly supported by the authors of the current guide). The emphasis of the current guide is on making good use of the GUM, on aspects that yield greater generality, and especially on the provision in some cases of measurement uncertainties that are more objectively based and numerically more sustainable. It is also concerned with validating the current usage of the GUM in circumstances where there is doubt concerning its applicability. Many laboratories and accreditation organizations have very considerable investment in the use of "Mainstream" GUM (i.e., as summarized in Clause 8 of the GUM). It is vital that this use continues in a consistent manner (certainly in circumstances where it remains appropriate); this guide conforms to this attitude. An important alternative, Monte Carlo Simulation, is presented as a numerical approach to be used when the conditions for Mainstream GUM to apply do not apply.

The relationship of this guide to the work being carried out by the Joint Committee on Guides in Metrology to revise the GUM is indicated.

An intention of this guide is to try to promote a scientific attitude to uncertainty evaluation rather than simply provide mechanistic procedures whose applicability is questionable in some circumstances.

# Contents

# Chapter 1

# Scope

## 1.1 Software Support for Metrology Programme

Almost all areas of metrology increasingly depend on software. Software is used in data acquisition, data analysis, modelling of physical processes, data visualisation, presentation of measurement and calibration results, evaluation of uncertainties, and information systems for the effective management of laboratories. The accuracy, interpretation and reporting of measurements all depend on the correctness of the software used. The UK's *Software Support for Metrology* (SS*f*M) Programme is designed to tackle a wide range of generic issues associated with mathematics, statistics, numerical computation and software engineering in metrology. The first Programme, spanning the period April 1998 to March 2001, is organised into four themes:

**Modelling techniques:** Modelling discrete and continuous data, uncertainties and statistical modelling, visual modelling and data visualisation, data fusion,

**Validation and testing:** Testing spreadsheet models and other packages used in metrology, model validation, measurement system validation, validation of simulated instruments,

**Metrology software development techniques:** Guidance on the development of software for metrology, software re-use libraries, mixed language programming and legacy software, development of virtual instruments,

**Support for measurement and calibration processes:** The automation of measurement and calibration processes, format standards for measurement data.

There are two further strands of activity concerning i) the assessment of the status of mathematics and software in the various metrology areas and ii) technology transfer.

The overall objective of the programme is the development and promotion of best practice in mathematical and computational disciplines throughout metrology through the publication of reports, case studies and best practice guides and organisation of seminars, workshops and training courses. An overview of the SS*f*M programme is available [48, 50].

This document is a deliverable associated with the first theme, modelling techniques, specifically that part of the theme concerned with uncertainties and statistical modelling.

## 1.2 Structure of the Guide

In summary, this best-practice guide provides information relating to

1. The use of statistical modelling to aid the construction of an *input-output model* as used in the Guide to the Expression of Uncertainty in Measurement (GUM) [1] (Section 2.1)

2. The objective of uncertainty evaluation (Section 2.2)

3. A statement of the main problem addressed in the area of uncertainty evaluation (Section 3.1)

4. The main stages of uncertainty evaluation, including a generally applicable two-phase procedure (Section 4)

5. Procedures for uncertainty evaluation and particularly for determining a coverage interval for the measurand (Section 5)

6. A classification of the main model types and guidance on the application of Mainstream GUM to these models (Section 6)

7. Details of a general numerical procedure, Monte Carlo Simulation, for uncertainty evaluation (Section 7)

8. A facility that enables the results of Mainstream GUM to be validated, thus providing assurance that Mainstream GUM can legitimately continue to be used in appropriate circumstances (Section 8)

9. Examples to illustrate the various aspects of this guide (Section 9).

## 1.3 Summary

This guide provides best practice on the evaluation of uncertainties within metrology and the support to this discipline given by statistical modelling. Central to considerations is a measurement system or process, having input quantities that are (invariably) inexact, and an output quantity that consequently is also inexact. The input quantities represent measurements or other information obtained from manufacturers' specifications and calibration certificates. The output represents a well-defined physical quantity to be measured, the *measurand*.[1] The objective of uncertainty evaluation is to model the system, including the quantification of the inputs to it, accounting for the nature of their inexactness, and to determine the model output, quantifying the extent and nature of

[1] In some instances the output quantities may not individually have physically meaning. An example is the set of coefficients in a polynomial representations of a calibration curve. Together, however, the set of quantities (coefficients) define a physically meaningful entity, the calibration curve.

*its* exactness.[2] A main requirement is to ascribe to the measurand a so-called *coverage interval* that contains the result of measurement, the "best estimate" of the model output quantity, and that can be expected to include a specified proportion, e.g., 95%, of the distribution of values that could reasonably be attributed to the measurand.[3]

A key document in the area of uncertainty evaluation is the Guide to the Expression of Uncertainty in Measurement (GUM) [1]. The GUM provides a "Mainstream" procedure[4] for evaluating uncertainties that has been adopted by many bodies. This procedure is based on representing the model input quantities in terms of estimated values and "standard uncertainties" that measure the dispersions of these values. These values and the corresponding uncertainties are "propagated" through (a linearized version of) the model to provide an estimate of the output quantity and its uncertainty. A means for obtaining a coverage interval for the measurand is provided. The procedure also accounts for the correlation effects that arise if the model input quantities are statistically interdependent.

In order to make the GUM more immediately applicable to a wider range of problems, a classification of model types is provided in this guide. The classification is based on

1. Whether there is one or more than one output quantity,

2. Whether the model or the quantities within it are real- or complex-valued, the latter arising in electrical, acoustical and optical metrology,

3. Whether the model is explicit or implicit, viz., whether or not it is possible to express the output quantity as a direct calculation involving the input quantities, or whether some indirect, e.g., iterative process, is necessitated.

Guidance on uncertainty evaluation based on Mainstream GUM principles is provided for each model type within the classification.

The model employed in the GUM is an *input-output model*, i.e., it expresses the measurand in terms of the input quantities. For relatively simple measurements, this form can straightforwardly be obtained. In other cases, this form does not arise immediately, and must be derived. Consideration is therefore given to *statistical modelling*, a process that relates the measurement data to the required measurement results and the errors in the various input quantities concerned. This form of modelling can then be translated into the "GUM model", in which the errors become subsumed in the input quantities and their influences summarized by uncertainties. Statistical modelling also covers the analysis and assignment of the nature of the inexactness of the model input quantities.

Although the GUM as a whole is a very rich document, there is much evidence that *Mainstream* GUM is the approach that is adopted by most practitioners. It is therefore vital that the fitness for purpose of this approach (and of any other approach) is assessed, generally and in individual applications. There

[2] Model validation, viz., the process of ascertaining the extent to which the model is adequate, is not treated in this guide. Detailed information on model validation is available [9].

[3] There may be more than one output, in which case a coverage *region* is required.

[4] Mainstream GUM is summarized in GUM Clause 8 and Section 6 of this guide.

are some limitations and assumptions inherent in the Mainstream GUM procedure and there are applications in metrology in which users of the GUM are unclear whether the limitations apply or the assumptions can be expected to hold in their circumstances. In such situations "other analytical or numerical methods" can be used, as is stated in the GUM (in Clause G.1.5). This "general" approach is contrasted in this guide with the mainstream approach. In particular, the limitations and assumptions at the basis of the "easy-to-use" formula inherent in Mainstream GUM are highlighted.

The GUM (through Clause G.1.5) does permit the practitioner to employ alternative techniques whilst remaining "GUM-compliant". However, if such techniques are to be used they must have certain credentials in order to permit them to be applied in a sensible way. Part of this guide is concerned with these techniques, their properties and their credentials.

It is natural, in examining the credentials of any alternative scientific approach, to re-visit established techniques to confirm or otherwise their appropriateness. In that sense it is appropriate to re-examine the principles of Mainstream GUM to discern whether they are fit for purpose. This task is not possible as a single "general health check". The reason is that there are circumstances when the principles of Mainstream GUM cannot be bettered by any other candidate technique, but there are others when the quality of the Mainstream GUM approach is not quantified. The circumstances in which Mainstream GUM is unsurpassed are when the model relating the input quantities $X_1, \ldots, X_n$ to the measurand $Y$ is *additive*, viz.,

$$Y = a_1 X_1 + \cdots + a_n X_n,$$

for any constants $a_1, \ldots, a_n$, any value of $n$, however large or small, and when the input quantities $X_i$ have independent Gaussian distributions. In other circumstances, Mainstream GUM generally provides an *approximate* solution: the quality of the approximation depends on the model and its input quantities and the magnitudes of their uncertainties. The approximation may in many cases be perfectly acceptable for practical application. In some circumstances this may not be so. See the statement in Clause G.6.6 of the GUM.

The concept of a *model* remains central to these alternative approaches. This guide advocates the use of such an alternative approach in circumstances where there is doubt concerning the applicability of Mainstream GUM. Guidance is provided for this approach. The approach is *numerical*, being based on Monte Carlo Simulation. It is thus computationally intensive, but nevertheless the calculation times taken are often only seconds or sometimes minutes on a PC, unless the model is especially complicated.

It is shown how the alternative approach can also be used to *validate* Mainstream GUM and thus in any specific application confirm (or otherwise) that this use of the GUM is *fit for purpose*, a central requirement of the Quality Management Systems operated by many organizations. In instances where the approach indicates that the use of Mainstream GUM is invalid, the approach can itself subsequently be used for uncertainty evaluation, in place of Mainstream GUM, in that it is consistent with the *general* principles (Clause G.1.5) of the GUM.

An overall attitude taken to uncertainty evaluation in this guide is that it consists of two phases. The first phase, *formulation*, constitutes building the

model and quantifying statistically its inputs. The second phase, *calculation*, consists of using this information to determine the model output quantity and quantify it statistically.

The concepts presented are demonstrated by examples, some chosen to emphasize a particular point and others taken from particular areas of metrology. Each of these examples illustrates Mainstream GUM principles or the recommended alternative approach or both, including the use of the latter as a validation facility for the former.

An account is included of the current situation concerning the revision of the GUM, a process that is taking place under the auspices of the Joint Committee for Guides in Metrology (JCGM).[5] This revision is concerned with amplifying and emphasizing key aspects of the GUM in order to make the GUM more readily usable and more widely applicable. Any published revision to the GUM, at least in the immediate future, would make no explicit change to the existing document, but enhance its provisions by the addition of supplemental guides. The approaches to uncertainty evaluation presented here are consistent with the developments by the JCGM in this respect, as is the classification of model types given. This best-practice guide will be updated periodically to account for the work of this committee. It will also account for the work of standards committees concerned with various aspects of measurement uncertainty, awareness of requirements in the areas indicated by workshops, etc., organized within SS*f*M and elsewhere, and technical developments.

Two of the authors of this guide are members of the Working Group of the JCGM that is concerned with GUM revision and of other relevant national or international committees, including British Standards Committee Panel SS/6/-/3, Measurement Uncertainty, CEN/BT/WG 122, Uncertainty of Measurement, and ISO/TC 69/SC 6, Measurement Methods and Results.

It is assumed that users of this guide have reasonable familiarity with the GUM.

A companion document [20] provides specifications of relevant software for uncertainty evaluation when applying some of the principles considered here.

## 1.4 Acknowledgements

This guide constitutes part of the deliverable of Project 1.2, "Uncertainties and statistical modelling" within the UK Department of Industry's National Measurement System *Software Support for Metrology* Programme 1998–2001. It has benefited from many sources of information. These include

- SS*f*M workshops
- The Joint Committee for Guides in Metrology
- National and international standards committees
- Consultative Committees of the Comitè International des Poids et Mesures (CIPM)
- The (UK) Royal Statistical Society

[5] The Web address of the JCGM is http://www.bipm.fr/enus/2_Committees/JCGM.shtml.

- The National Engineering Laboratory
- National Measurement Institutes
- The United Kingdom Accreditation Service
- UK industry
- Conferences in the Advanced Mathematical and Computational Tools in Metrology series [12, 13, 14, 15, 16]
- Literature on uncertainties, statistics and statistical modelling
- Many individual contacts.[6]

[6]The contacts are far too numerous to mention. Any attempt to enumerate them would risk offence to those unintentionally omitted.

# Chapter 2

# Introduction

## 2.1 Uncertainty and statistical modelling

Measurements contain errors. When a quantity is measured, the actual value obtained is not the true value, but some value that departs from it to a greater or lesser extent—an approximation to it. If that quantity were to be measured a number of times, in the same way and in the same circumstances, a different value each time would in general be obtained.[1] These repeated measurements would form a "cluster", the "size" of which would depend on the nature and quality of the measurement process. The "centre" of the cluster would provide an estimate of the quantity that generally can be expected to be more reliable than individual measurements. The "size" of the cluster would provide quantitative information relating to the quality of this central value as an estimate of the quantity. It will not furnish all the information of this type, however. The measuring instrument is likely to provide values that are not scattered about the true value of the quantity, but about some other value offset from it.

Take the domestic bathroom scales. If they are not set such that the display reads zero when there is nobody on the scales, when used to weigh a person or an object the observed weight can be expected to be offset from what it should be. No matter how many times the person's weight is taken and averaged,[2] because the scatter of values would be centred on an offset value, the effect of this offset is inherently present in the result. A further effect is that scales possess "stiction", i.e., they do not necessarily return consistently to the "starting position" each time a person gets on and off.

There are thus two main effects, in this example and in general. The first is a "random" effect associated with the fact that when a measurement is repeated it will generally be different from the previous value. It is random in that there is no way to predict from previous measurements exactly what the next one would be.[3] The second effect is a systematic effect (a bias) associated with the fact that the measurements contain an offset.

In practice there can be a number of contributions to the random effect and

[1] This statement assumes that the recording device has sufficient resolution to distinguish between different values.

[2] There is a variety of ways of taking an average, but the choice made does not affect the argument.

[3] If a prediction *were* possible, allowance for the effect could be made!

to the systematic effect, both in this situation and in many other situations. Depending on the application, the random effect may dominate, the systematic effect may dominate or the effects may be comparable.

In order to make a statement concerning the measurement of the quantity of interest it is typically required to provide a value for the measurand and an associated "uncertainty". The value is (ideally) a "best estimate" of the measurand and the uncertainty a numerical measure of the quality of the estimate.

The above discussion concerns the measurement of a particular quantity. However, the quantity *actually* measured by the device or instrument used is rarely the result required in practice. For instance, the display on the bathroom scales does not correspond to the quantity measured. The raw measurement might be that of the extension of a spring in the scales whose length varies according to the load (the weight of the person on the scales).

The raw measurement is therefore *converted* or *transformed* into the required form, the *measurement result*. The latter is an estimate of the *measurand*, the physical quantity that is the subject of measurement.

For a perfect (linear) spring, the conversion is straightforward, being based on the fact that the required weight is proportional to the extension of the spring. The display on the scales constitutes a *graduation* or *calibration* of the device.

For a domestic mercury thermometer, the raw measurement is the height of a column of mercury. This height is converted into a temperature using another proportional relationship: a change in the height of the column is proportional to the change in temperature, again a calibration.

A relationship of types such as these constitutes a *rule* for converting the raw measurement into the measurement result.

In metrology, there are very many different types of measurement and therefore different rules. Even for one particular type of measurement there may well be more than one rule, perhaps a simple rule (e.g., a proportional rule) for everyday domestic use, and a sophisticated rule involving more complicated calculations (a *nonlinear* rule, perhaps) that is capable of delivering more accurate results for industrial or laboratory purposes.

Often, measurements are *repeated* and averaged in some way to obtain a more reliable result.

The situation is frequently more general in another way. There is often a number of *different* raw measurements that contribute to the estimation of the measurand. Here, the concern is not simply repeated measurements, but intrinsically different measurements, e.g., some temperature measurements and some displacement measurements. Also, there may be more than one measurand. For instance, by measuring the length of a bar at various temperatures it may be required to determine the coefficient of expansion of the material of which the bar is made and also to determine the length of the bar at a temperature at which it may not have been measured, e.g., 27 °C, when measurements were made at 20, 22, 24, 26, 28 and 30 °C.

In addition to raw data, representing measurements, there is another form of data that is also frequently fed into a rule in order to provide a measurement result. This additional data relates to a variety of "constants", each of which can be characterized as having a value and a distribution about it to represent the imperfect knowledge of the value. An example is a material constant such as modulus of elasticity, another is a calibrated dimension of an artefact such

as a length or diameter, and another is a *correction* arising from the fact that a measurement was made at, say, 22 °C rather than the stipulated 20 °C.

The complete set of data items that are required by the rule to enable a measurement result to be produced is known as the *input quantities*. The rule is usually referred to as a *model* because it is the use of physical modelling (or perhaps empirical modelling or both types of modelling) [22] of a measurement, measurement system or measurement process that enables the rule to be established. The model output quantities are used to estimate the measurands.

This guide is concerned with the problem of characterizing the nature of the errors in the estimates of the measurands given the model, the input quantities and information concerning the errors in these quantities. Some advice is given on assigning statistical properties to the input quantities. Because the form of the model varies enormously over different metrology disciplines, it is largely assumed that a (physical) model is available (having been derived by the experts in the appropriate area). The use of statistical modelling is considered, however, in the context of capturing the error structure of a problem. Model validity is not specifically addressed. Information is available in a companion publication [9].

In particular, this guide reviews several approaches to the problem, including the widely-accepted GUM approach. It reviews the interpretation of the GUM that is made by many organisations and practitioners concerned with measurements and their analysis and the presentation of measurement results (Section 6 of this guide).

The point is made that this interpretation is subject to *limitations* that are insufficiently widely recognized. These limitations have, however, been indicated [61] and are discussed in Section 5.4.

An approach free from these limitations is presented. It is a numerical method based on the use of *Monte Carlo Simulation* (MCS) and can be used

1. in its own right to characterise the nature of the inexactness in the measurement results,

2. to *validate* the approach based on the above-mentioned interpretation of the GUM.

MCS itself has deficiencies. They are of a different nature from those of Mainstream GUM, and to a considerable extent controllable. They are identified in Section 7.6.

The GUM does not refer *explicitly* to the use of MCS. However, this option was recognized during the drafting of the GUM. The ISO/IEC/OIML/BIPM draft (First Edition) of June 1992, produced by ISO/TAG 4/WG 3, states, as Clause G.1.5:

> If the relationship between $Y$ [the model output] and its input quantities is nonlinear, or if the values available for the parameters characterizing the probabilities of the $X_i$ [the inputs] (expectation, variance, higher moments) are only estimates and are themselves characterized by probability distributions, and a first order Taylor expansion is not an acceptable approximation, the distribution of $Y$ cannot be expressed as a convolution. In this case, numerical methods (such as Monte Carlo calculations) will generally be required and the evaluation is computationally more difficult.

In the published version of the GUM [1], this Clause had been modified to read:

> If the functional relationship between $Y$ and its input quantities is nonlinear and a first-order Taylor expansion is not an acceptable approximation (see 5.1.2 and 5.1.5), then the probability distribution of $Y$ cannot be obtained by convolving the distributions of the input quantities. In such cases, other analytical or numerical methods are required.

The interpretation made here of this re-wording is that "other analytical or numerical methods" cover any other *appropriate* approach.[4]

This interpretation is consistent with that of the National Institute of Standards and Technology of the United States [61]:

> [Clause 6.6] The NIST policy provides for exceptions as follows (see Appendix C):
>
> It is understood that any valid statistical method that is technically justified under the existing circumstances may be used to determine the equivalent of $u_i$ [the standard deviation of the $i$th input quantity], $u_c$ [the standard deviation of the output], or $U$ [the half-width of a coverage interval for the output, under a Gaussian assumption]. Further, it is recognised that international, national, or contractual agreements to which NIST is a party may occasionally require deviation from NIST policy. In both cases, the report of uncertainty must document what was done and why.

Further, within the context of statistical modelling in analyzing the homogeneity of reference materials, it is stated [39]:

> [Clause 9.2.3] ... where lack of a normal distribution is a problem, robust or non-parametric statistical procedures may be used to obtain a valid confidence interval for the quantity of interest.

This guide adheres to these broad views. The most important aspect relates to *traceability of the results* of an uncertainty evaluation. An uncertainty evaluation should include

1. all relevant information relating to the model and its input quantities,
2. an estimate of the measurand and an associated coverage interval (or coverage region),
3. the manner in which these results were determined, including all assumptions made.

There would also appear to be valuable and relevant interpretations and considerations in the German standard DIN 1319 [27]. An official English-language translation of this standard would not seem to be available.

There has been a massive investment in the use of the GUM. It is essential that this investment is respected and that this guide is not seen as deterring

[4] That this interpretation is correct has been confirmed by JCGM/WG1.

the continuation of its use, at least in circumstances where such usage can be demonstrated to be appropriate.

In this respect, a *recommended* validation procedure for Mainstream GUM is provided in this guide. The attitude taken is that if the procedure demonstrates in any particular circumstance that this usage is indeed valid, the Mainstream GUM procedure can legitimately continue to be used in that circumstance. The results of the validation can be used to record the fact that fitness for purpose in this regard has been demonstrated. If the procedure indicates that there is doubt concerning the validity of Mainstream GUM usage, there is a case for investigation. Since in the latter case the recommended procedure forms a constituent part (in fact the major part) of the validation procedure, this procedure can be used in place of the Mainstream GUM approach. Such use of an alternative procedure is consistent with the broader principles of the GUM (Section 5 of this guide and above).

There is another vital issue facing the metrologist. In a measurement situation it is necessary to characterise the nature of the errors in the input quantities and to develop the model for the measurand in terms of these quantities. Carrying out these tasks can be far from easy. Some advice is given in this regard. However, written advice can only be general, although examples and case studies can assist. In any one circumstance, the metrologist has the responsibility, perhaps with input from a mathematician or statistician if appropriate, of characterizing the input quantities and building the model.

The Mainstream GUM procedure and the recommended approach using Monte Carlo Simulation both utilize this information (but in different ways). As mentioned, the former possesses some limitations that the latter sets out to overcome (but again see Section 7.7 concerning implementation).

There is often doubt concerning the nature of the errors in the input quantities. Are they Gaussian or uniform, or do they follow some other distribution? The metrologist needs to exercise best judgement in this regard (see Section 4.3). Even then there may be aspects that cannot fully be quantified.

It is regarded as important that this incomplete lack of knowledge, that can arise in various circumstances, is handled by *repeating* the exercise of characterizing the errors in the outputs. By this statement it is meant that any assumptions relating to the nature of the errors in the inputs are changed to other assumptions that in terms of the available knowledge are equally valid. Similarly, the information that leads to the model may be incomplete and therefore changes to the model consistent with this lack of knowledge made.

The sensitivity of the output quantities can then be assessed as a function of such perturbations by repeating the evaluation.

The attitude here is that whatever the nature of the input quantities and the model, even (and especially!) if some subjective decisions are made in their derivation, the nature of the outputs should then follow objectively and without qualification *from this information*, rather than in a manner that is subject to limitations, in the form of effects that are difficult to quantify and *beyond the control of the practitioner*.

In summary, the attitude that is generally promoted in this guide is that as far as economically possible use should be made of all available knowledge. In particular, (a) the available knowledge of the input quantities should be embodied within their specification, (b) a model that relates these input quantities to the measurand should carefully be constructed, and (c) the calculation of

uncertainty should be carried out in terms of this information.

## 2.2 The objective of uncertainty evaluation

Uncertainty evaluation is the generic term used in this guide to relate to any aspect of quantifying the extent of the inexactness in the outputs of a model to inexactness in the model input quantities. Also, the model itself may be inexact. If that is the case, the nature and extent of the inexactness also need to be quantified and its influence on the output quantities established. The inexactness of the model outputs is also influenced by any algorithm or software that is used to determine the output quantity given the input quantities. Such software may incorporate *approximate* algorithmic techniques that impart an additional uncertainty.

**Example 1** *Approximate area under a calibration curve*

Consider a model necessitating the determination of an integral representing the area under a calibration curve. An algorithm might utilize the trapezoidal or some other approximate numerical quadrature rule. Numerical errors will be committed in the use of this rule. They depend on the spacing of the ordinates used and on the extent of the departure of the curve from linearity. The consequent uncertainties would need to be evaluated.

The uncertainty evaluation process could be at any level required, depending on the application. At one extreme it could involve determining the standard deviation of an estimate of the measurand for a simple model having a single output. At the other extreme it might be necessary to determine the joint probability distribution of a set of output quantities of a complicated complex-valued model exhibiting non-Gaussian behaviour, and from that deduce a coverage region for the vector of measurands at a stipulated level of probability.

The objective of uncertainty evaluation can be stated as follows:

*Derive (if not already available) a model relating a set of measurands to (input) quantities (raw measurements, suppliers' specifications, etc.) that influence them. Establish the statistical properties of these input quantities. Calculate (in a sense required by context) estimates of the measurands and the uncertainty of these estimates.*

A mathematical form for this definition is given in Section 3.1.

This objective may in its context be well defined or not. In a case where it is well defined there can be little dispute concerning the nature of the results, presuming they have been calculated correctly. If it is not well defined, it will be necessary to augment the information available by assumptions or assertions in order to establish a well-defined problem. It will be necessary to ensure that the assumptions and assertions made are as sensible as reasonably possible in the context of the application. It will equally be necessary to make the assumptions and assertions overt *and to record them*, so that the results can be reproduced and defended, and perhaps subsequently improved.

In very many cases the objective of uncertainty evaluation will be to determine a coverage interval (or coverage region) for the measurand. Commonly, this coverage interval will be at the 95% level of probability. There is no compelling

*scientific* reason for this choice. It almost certainly stems from the traditional use of 95% in statistical hypothesis testing [11], although the reasons for the choice in that area are very different. The overriding reason for the use of 95% in uncertainty evaluation is a practical one. It has become so well established that for purpose of comparison with other results its use is almost mandated. Another strong reason for the use of 95% is the considerable influence of the Mutual Recognition Arrangement concerning the comparison of national measurement standards and of calibration and measurement certificates issued by national metrology institutes [7]. See Appendix A.4 for a discussion.

Such an interval will be referred to in this guide as a *95% coverage interval*.

It can be argued that if a coverage interval at some other level of probability is quoted, it can be "converted" into one at some other level. Indeed, a similar operation is recommended in the GUM, when information concerning an input distribution is converted into a standard deviation (standard uncertainty in GUM parlance). The standard deviations together with sensitivity coefficients are combined to produce the standard deviation of the output, from which a coverage interval is obtained by multiplication by a factor. The factor is selected based on the *assumption* that the output distribution is Gaussian.

That this process gives rise to difficulties in some cases can be illustrated using a simple example. Pre-empting the subsequent discussion, consider the model $Y = X_1 + X_2 + \ldots$, where $X_1, X_2, \ldots$ are the input quantities and $Y$ the output quantity. Assume that all terms but $X_1$ have a small effect, and $X_1$ has a uniform distribution. The above-mentioned GUM procedure gives a 95% coverage interval for $Y$ that is longer than the 100% coverage interval for $X_1$!

Instances of this type would appear to be not uncommon. For instance, the EA guide [28] gives three examples arising in the calibration area.

This possibility is recognised by the GUM:

> [GUM Clause G.6.5] ... Such cases must be dealt with on an individual basis but are often amenable to an analytic treatment (involving, for example, the convolution of a normal distribution with a rectangular distribution ...

The statement that such cases must be dealt with on an individual basis would appear to be somewhat extreme. Indeed, such a treatment is possible (cf. Sections 5.1 and 5.1.2), but is not necessary, since Monte Carlo Simulation (Section 7) generally operates effectively in cases of this type.

The interpretation [63] of the GUM by the United Kingdom Accreditation Service recommends the inclusion of a dominant uncertainty contribution by adding the term linearly to the remaining terms combined in quadrature. This interpretation gives rise generally to a more valid result, but remains an approximation. The EA Guide [28] provides some analysis in some such cases.

It is emphasized that a result produced according to a fixed recipe that is not universally applicable, such as Mainstream GUM, may well be only *approximately* correct, and the degree of approximation difficult to establish.

The concern in this guide is with reliable uncertainty evaluation, in that the results will not exhibit inconsistent or anomalous behaviour, however simple or complicated the model may be.

Appendix A reviews some relevant statistical concepts.

## 2.3 Standard deviations and coverage intervals

The most important statistic to a metrologist is a *coverage interval* corresponding to a specified probability, e.g., an interval that is expected to contain 95% of the values that could be attributed to the measurand. This interval is the 95% coverage interval considered above.

There is an important distinction between the nature of the information needed to determine the standard deviation of the estimate of the output quantity and a coverage interval for the measurand.

The mean and standard deviation can be determined knowing the distribution of the output. The converse is not true.

**Example 2** *Deducing a mean and standard deviation from a distribution, but not the converse*

As an extreme example, consider a random variable $X$ that can take only two values, $a$ and $b$, with equal probability. The mean is $\mu = (a + b)/2$ and the standard deviation $\sigma = |b - a|/2$. However, given only the values of $\mu$ and $\sigma$, there is no way of deducing the distribution. If a Gaussian distribution were assumed, it would be concluded that the interval $\mu \pm 1.96\sigma$ contained 95% of the distribution. In fact, the interval contains 100% of the distribution, as does the interval $\mu \pm \sigma$, of about half that length.

Related comments are made in Clause G.6.1 of the GUM. Although knowledge of the mean and standard deviation is valuable information, without *further* information it conveys nothing about the *manner* in which the values are distributed.[5] If, however, it is known that the underlying distribution is Gaussian, the distribution of the output quantity is completely described since just the mean and standard deviation fully describe a Gaussian distribution. A similar comment can be made for some other distributions. Some distributions require additional parameters to describe them. For instance, in addition to the mean and standard deviation, a Student's-$t$ distribution requires the number of degrees of freedom to specify it.

Thus, if the form of the distribution is known, from analysis, empirically or from other considerations, the determination of an appropriate number of *statistical parameters* will permit it to be quantified. Once the quantified form of the distribution is available, it is possible to calculate a *percentile*, i.e., a value for the measurement result such that, according to the distribution, the corresponding percentage of the possible values of the measurement result is smaller than that value. For instance, if the 25-percentile is determined, 25% of the possible values can be expected to lie below it (and hence 75% above it). Consider the determination of the 2.5-percentile and the 97.5-percentile. 2.5% of the values will lie to the left of the 2.5-percentile and 2.5% to the right of the 97.5-percentile. Thus, 95% of the possible values of the measurement result lie between these two percentiles. These points thus constitute the endpoints of a 95% coverage interval for the measurand.

The 2.5-percentile of a distribution can be thought of as a point a certain number of standard deviations below the mean and the 97.5-percentile as a point a certain number of standard deviations above the mean. The numbers of

[5] See, however, the maximum entropy considerations in Appendix D.2.

standard deviations to be taken depends on the distribution. They are known as *coverage factors*. They also depend on the coverage interval required, 90%, 95%, 99.8% or whatever.

For the Gaussian distribution and the Student's-$t$ distribution, the effort involved in determining the numbers of standard deviations to be taken has been embodied in tables and software functions.[6] Since these distributions are symmetric about their means, the coverage factors for pairs of percentiles that sum to 100, such as the above 2.5- and 97.5-percentiles, are identical. This statement is not generally true for asymmetric probability distributions.

In order to determine percentiles in general, it is necessary to be able to evaluate the inverse $G^{-1}$ of the distribution function $G$ (Section A.3 in Appendix A). For well-known distributions, such as Gaussian and Student's-$t$, software is available in many statistical and other libraries for this purpose. Otherwise, values of $x_p = G^{-1}(p)$ can be determined by using a *zero finder* to solve the equation $G(x_p) = p$ (cf. Section 6.2.3).

The coverage interval is not unique, even in the symmetric case. Suppose that a probability density function (Appendix A) $g(x) = G'(x)$ is unimodal (single-peaked), and that a value of $\alpha$, $0 < \alpha < 1$, is given. Consider any interval $[a,\ b]$ that satisfies

$$G(b) - G(a) = \int_a^b g(x)dx = 1 - \alpha.$$

Then [52],

1. $[a,\ b]$ is a $100(1-\alpha)\%$ coverage interval. For instance, if $a$ and $b$ are such that
$$G(b) - G(a) = 0.95,$$
95% of the possible values of $x$ lie between $a$ and $b$,

2. The shortest such interval is given by $g(a) = g(b)$. $a$ lies to the left of the mode (the value of $x$ at which $g(x)$ is greatest) and $b$ to the right,

3. If $g(x)$ is symmetric, not only is the shortest such interval given by $g(a) = g(b)$, but also $a$ and $b$ are equidistant from the mode.

[6] In most interpretations of the GUM (Mainstream GUM), the model output is taken as Gaussian or Student's-$t$.

# Chapter 3

# Uncertainty evaluation

## 3.1 The problem formulated

As discussed in Section 2.1, regardless of the field of application, the physical quantity of concern, the *measurand*, can rarely be measured directly. Rather, it is determined from a number of contributions, or input quantities, that are themselves measurements or derived from other measurements or information.

The fundamental relationship between the input quantities and the measurand is the model. The input quantities, $n$, say, in number, to the model are denoted by $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ and the measurand, the output quantity, by $Y$.[1] The model

$$Y = f(\mathbf{X}) = f(X_1, \ldots, X_n)$$

can be a mathematical formula, a step-by-step calculation procedure, computer software or other prescription. Figure 3.1 shows an input-output model to illustrate the "propagation of uncertainties" (GUM [1]). The model has three input quantities $X = (X_1, X_2, X_3)^{\mathrm{T}}$, where $X_i$ is estimated by a value $x_i$ with standard deviation $u(x_i)$. It has a single output $Y \equiv Y_1$, with estimated value $y = y_1$ and standard deviation $u(y) = u(y_1)$. The estimate of the measurand is termed the *measurement result*. In a more complicated circumstance, the errors in the input quantities would be interrelated, i.e., correlated, and additional information would be needed to quantify the correlations.

There may be more than one output quantity, viz., $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\mathrm{T}}$. In this case the model is

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}) = \mathbf{f}(X_1, \ldots, X_n),$$

where $\mathbf{f}(\mathbf{X}) = (f_1(\mathbf{X}), \ldots, f_m(\mathbf{X}))$, a vector of model functions. In full, this "vector model" is

$$\begin{aligned} Y_1 &= f_1(X_1, \ldots, X_n), \\ Y_2 &= f_2(X_1, \ldots, X_n), \\ &\vdots \\ Y_m &= f_m(X_1, \ldots, X_n). \end{aligned}$$

[1] A single input quantity (when $n = 1$) will sometimes be denoted by $X$ (rather than $X_1$).

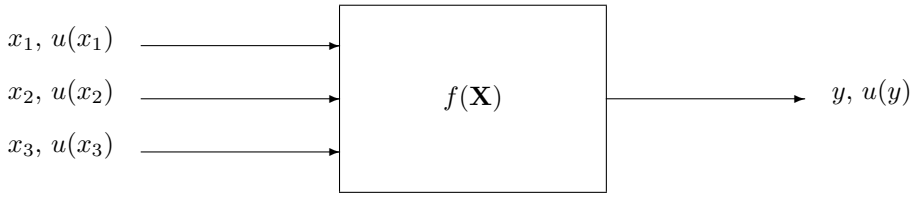


Figure 3.1: Input-output model illustrating the propagation of uncertainties. The model has three input quantities $X = (X_1, X_2, X_3)^{\mathrm{T}}$, estimated by a value $x_i$ with standard deviation $u(x_i)$, for $i = 1, 2, 3$. There is a single measurand (output quantity) $Y \equiv Y_1$, estimated by the measurement result $y$, with standard deviation $u(y)$.

The output quantities $\mathbf{Y}$ would almost invariably be correlated in this case, since in general each output quantity $Y_j$, $j = 1, \ldots, m$, would depend on several or all of the input quantities.

A model with a single output quantity $Y$ is known as a *univariate model*. A model with $m$ ($> 1$) output quantities $\mathbf{Y}$ is known as a *multivariate model*.

In statistical parlance all input quantities $X_i$ are regarded as *random variables*, regardless of their source (cf. [65]). The model function $f(X_1, \ldots, X_n)$ is an *estimator*. The output quantity $Y$ is also a random variable. Realizations $x_i$ of the $X_i$ are *estimates* of the inputs. $f(x_1, \ldots, x_n)$ provides an estimate of the output quantity, the measurand.[2] A comparable statement applies to a multivariate model.

## 3.2 The two phases of uncertainty evaluation

Uncertainty evaluation consists of two phases, *formulation* and *calculation*. In Phase 1 the metrologist derives the model, perhaps in collaboration with a mathematician or statistician. The metrologist also provides the model inputs, qualitatively, in terms of their probability density functions (pdf's) (uniform, Gaussian, etc.) (Appendix A.2), and quantitatively, in terms of the parameters of these functions (e.g., central value and semi-width for a uniform pdf, or mean and standard deviation for a Gaussian pdf), including correlation parameters for joint pdf's. These pdf's are obtained from an analysis of series of observations

[2]This estimator may be biased. The mean of the probability density function of the output quantity is unbiased. It is expected that the bias will be negligible in many cases. The bias results from the fact that the value of $Y$ obtained by evaluating the model at the input estimates $\mathbf{x}$ is not in general equal to the value of $Y$ given by the mean of the pdf $g(y)$ of $Y$. These values will be equal when the model is linear in $\mathbf{X}$, and close if the model is mildly nonlinear or if the uncertainties of the input quantities are small. See Section 3.2. A demonstration of the bias is given by the simple model $Y = X^2$, where $X$ is assigned a Gaussian distribution with mean zero and standard deviation $u$. The mean of the pdf describing the input quantity $X$ is zero. The corresponding value of the output quantity $Y$ is also zero. However, the mean value of the pdf describing $Y$ cannot be zero, since $Y \geq 0$, with equality occurring only when $X = 0$. (This pdf is in fact a $\chi^2$ distribution with one degree of freedom.)

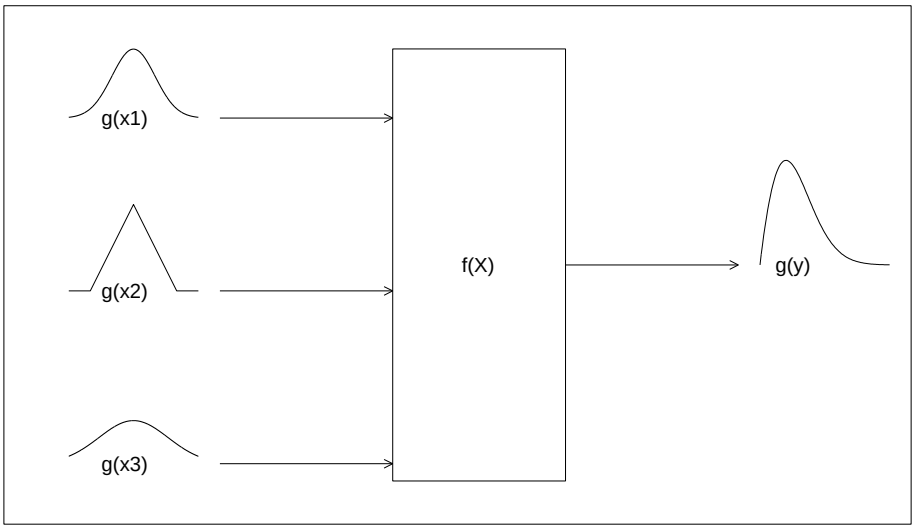


Figure 3.2: Input-output model illustrating the propagation of distributions. The model has three input quantities $X = (X_1, X_2, X_3)^{\mathrm{T}}$, where $X_1$ has a Gaussian pdf $g_1(x_1)$, $X_2$ a triangular pdf $g_2(x_2)$ and $X_3$ a (different) Gaussian pdf $g_1(x_3)$. The single output quantity $Y \equiv Y_1$ is illustrated as being asymmetric, as can arise for nonlinear models where one or more of the input pdf's has a large standard deviation.

[1, Clauses 2.3.2, 3.3.5] or are based on scientific judgement using all the relevant information available [1, Clauses 2.3.3, 3.3.5], [61].

In the case of uncorrelated input quantities and a single output quantity, Phase 2 of uncertainty evaluation is as follows. Given the model $Y = f(\mathbf{X})$, where $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$, and the pdf's $g_i(x_i)$ (or the distribution functions $G_i(x_i)$) of the input quantities $X_i$, for $i = 1, \ldots, n$, determine the pdf $g(y)$ (or the distribution function $G(y)$) for the measurand $Y$.

The mean of $g(y)$ is an unbiased estimate of the output quantity.

Figure 3.2 shows the counterpart of Figure 3.1 in which the pdf's (or the corresponding distribution functions) of the input quantities are propagated through the model to provide the pdf (or distribution function) of the measurand.

It is reiterated that once the pdf (or distribution function) of the measurand $Y$ has been obtained, any statistical information relating to $Y$ can be produced from it. In particular, a coverage interval for the measurand at a stated level of probability can be obtained.

When the input quantities are correlated, in place of the $n$ individual pdf's $g_i(x_i)$ $i = 1, \ldots, n$, there is a joint pdf $\mathbf{g}(\mathbf{x})$. An example of a joint pdf is the multivariate Gaussian pdf (Section 4.3.2). In practice this joint pdf may be decomposable. For instance, in some branches of electrical, acoustical and optical metrology, the input quantities may be complex-valued. The real and imaginary parts of each such variable are generally correlated and thus each has an associated $2 \times 2$ covariance matrix. See Section 6.2.5. Otherwise, the input quantities may or may not be correlated.

If there is more than one output quantity, $\mathbf{Y}$, these outputs will almost invariably need to be described by a joint pdf $\mathbf{g}(\mathbf{y})$, since each output quantity generally depends on all the input quantities. See Section 9.7 for an important exception.

Phase 2, calculation, involves the derivation of the measurement result and the associated uncertainty, given the information provided by Phase 1. It is computational and requires no further information from the metrology application. The uncertainty is commonly provided as a coverage interval. A coverage interval can be determined once the distribution function $G(y)$ (Appendix A) has been derived. The endpoints of a 95% coverage interval[3] are given (Section 2.3) by the 0.025- and 0.975-quantiles of $G(y)$, the $\alpha$-quantile being the value of $y$ such that $G(y) = \alpha$.[4]

It is usually sufficient to quote the uncertainty of the measurement result to one or at most two significant figures. See Section 7.7.1. In general, further figures would be spurious, because the information provided in Phase 1 is typically imprecise, involving a number of estimates and assumptions. The attitude taken here though is that the second phase should not exacerbate the consequences of the decisions made in Phase 1.[5]

The pdf $g(y)$ pertaining to $Y$ cannot generally be expressed in simple or even closed mathematical form. Formally, if $\delta(\cdot)$ denotes the Dirac delta function,

$$g(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x})\delta(y - f(\mathbf{x}))dx_n dx_{n-1} \ldots dx_1 \qquad (3.1)$$

[19]. Approaches for determining $g(y)$ or $G(y)$ are addressed in Section 5. That several approaches exist is a consequence of the fact that the determination of $g(y)$ and/or $G(y)$ ranges from being very simple to extremely difficult, depending on the complexity of the model and its input pdf's.

[3] 95% coverage intervals are used in this guide, but the treatment applies more generally.

[4] There are many intervals having a coverage probability of 0.95, a general interval being given by the $\beta$- and $(0.95 + \beta)$-quantiles of $G(y)$, with $0 \leq \beta \leq 0.05$. The choice $\beta = 0.025$ is natural for a $G(y)$ corresponding to a symmetric pdf $g(y)$. It also has the shortest length for a symmetric pdf and, in fact, for a *unimodal* pdf (Section 2.3).

[5] This attitude compares with that in mathematical physics where a model (e.g., a partial differential equation) is constructed and then solved. The construction involves idealizations and inexact values for dimensional quantities and material constants, for instance. The solution process involves the application of hopefully sensible and stable methods in order to make some supported statements about the quality of the solution obtained to the *posed* problem.

# Chapter 4

# The main stages in uncertainty evaluation

In this guide, uncertainty evaluation is regarded as consisting of the two main phases indicated in Chapter 3.1.

Phase 1, formulation, consists of providing the model and quantifying the pdf's of the model input quantities. The constituent parts of Phase 1 are the three steps

1. Statistical modelling.
2. Input-output modelling.
3. Assignment of probability density functions to the input quantities.

Phase 2, calculation, consists of deriving the measurement result and its uncertainty, given the information provided by Phase 1. The constituent parts of Phase 2 are the two steps

1. Determination of the probability density function of the output quantity.
2. Provision of a coverage interval for the measurand.

## 4.1 Statistical modelling

Statistical modelling can be beneficial when a model is complicated, but is not always needed for simpler models. It is concerned with developing the relationships between the measurements made, the measurands and the errors of measurement.

**Example 3** *Straight-line calibration*

A common example of statistical modelling arises when fitting a calibration curve to data, representing, say, the manner in which displacement varies with temperature. The input quantities consist, for $i = 1, \ldots, n$, say, of a measured value $x_i$ of a response variable $x$ corresponding to the value $t_i$ of a control or independent variable $t$.

Consider a situation in which the $t_i$ are known with negligible error and $x_i$ has error $e_i$. Suppose that the nature of the calibration is such that a straight-line calibration curve $y = a_1 + a_2 x$ provides an adequate realization. Then, as part of the statistical modelling process [22], the equations

$$x_i = a_1 + a_2 t_i + e_i, \qquad i = 1, \ldots, n, \tag{4.1}$$

relate the observations $x_i$ to the calibration parameters $a_1$ (intercept) and $a_2$ (gradient) of the line and the errors $e_i$ in the measurements.

In order to establish values for $a_1$ and $a_2$ it is necessary to make an appropriate assumption about the nature of the errors $e_i$ [22]. For instance, if they can be regarded as realizations of independent, identically distributed Gaussian variables, the best unbiased estimate of the parameters is given by *least squares*. Specifically, $a_1$ and $a_2$ are given by minimizing the sum of the squares of the deviations $x_i - a_1 - a_2 t_i$, over $i = 1, \ldots, n$, with respect to $a_1$ and $a_2$, viz.,

$$\min_{a_1, a_2} \sum_{i=1}^{n} (x_i - a_1 - a_2 t_i)^2.$$

The model equations (4.1), with the solution criterion (least-squares), constitute the results of the statistical modelling process for this example.

There may be additional criteria. For instance, a calibration line with a negative gradient may make no sense in a situation where the gradient represents a physical parameter whose "true" value must always be greater than or equal to zero (or some other specified constant value). The overall criterion in this case would be to minimize the above sum of squares with respect to $a_1$ and $a_2$, as before, with the condition that $a_2 \geq 0$. This problem is an example of a *constrained* least-squares problem, for which sound algorithms exist [22]. In this simple case, however, the problem can be solved more easily for the parameters, but the uncertainties associated with $a_1$ and $a_2$ require special consideration. See the example in Section 9.6.



Appendix B discusses some of the modelling issues associated with the statistical modelling approach of this section and the input-output modelling approach of the next.

## 4.2 Input-output modelling

Input-output modelling is the determination of the model required by the GUM, also termed here the *GUM model*, in its approach to uncertainty evaluation and, as indicated in Section 3.1, also in this guide. An input-output model can be derived from a statistical model or formed directly.

In the GUM [1] a measurement system is modelled, as in Section 3.1, by a functional relationship between measured or influence quantities (the input quantities) $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ and the measurand (the output quantity) $Y$ in the form

$$Y = f(\mathbf{X}). \tag{4.2}$$

In practice this functional relationship does not apply directly to all measurement systems encountered, but may instead (a) take the form of an *implicit* relationship, $h(Y, \mathbf{X}) = 0$, (b) involve a *number* of measurands $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\mathrm{T}}$,

or (c) involve *complex-valued* quantities. Section 6 is concerned with the manner in which each model type within this classification can be treated within a "GUM" setting. Here, the concern is with the basic form (4.2).

**Example 4** *How long is a piece of string?*

The problem of establishing a simple model for the length of a piece of string, when measured with a tape, is considered (cf. [5]). The measurand is the length of the string. As part of Phase 1, formulation, a measurement model for string length is established. It depends on several input quantities. This model is expressed here as the sum of four terms. Each of these terms, apart from the first, is itself expressed as a sum of terms.[1] The model takes the form[2]

| | | |
|---|---|---|
| String length | = | Measured string length (1) |
| | + | Tape length error (2) |
| | + | String length error (3) |
| | + | Measurement process error (4), |

where

| | | |
|---|---|---|
| (1) Measured string length | = | Average of a number of repeated measurements |
| (2) Tape length error | = | Length error due to tape calibration imperfections |
| | + | Extension in tape due to stretching (negative if there is shrinking rather than stretching) |
| | + | Reduction in effective length of tape due to bending of the tape |
| (3) String length error | = | Reduction in effective string length due to string departing from a straight line |
| | + | Reduction in string length as a result of shrinking (negative if there is stretching rather than shrinkng) |
| (4) Measurement process error | = | Length error due to inability to align end of tape with end of string due to fraying of the string ends |
| | + | Length error due to the tape and the string not being parallel |
| | + | Error committed in assigning a numerical value to the measurement indicated by the tape |
| | + | Length error arising from the statistics of averaging a finite number of repeated measurements. |

Once this model is in place statements can be made about the nature of the various terms in the model as part of the first phase of uncertainty evaluation.

[1] The model can therefore be viewed as a multi-stage model (Section 4.2.3, although of course by substitution it can be expressed as a single model.

[2] In this formula, the error terms are to be expressed in a way that ensures each contribution has the correct numerical (±) sign.

Phase 2 can then be carried out to calculate the required uncertainty in the measured length.

There may be some statistical modelling issues in assigning pdf's to the input quantities. For instance, a Student's-$t$ distribution would probably be assigned to the measured string length (1), based on the mean and standard deviation of the repeated measurements, with a number of degrees of freedom one less than the number of measurements. As another instance, for tape bending, this effect can be approximated by a $\chi^2$ variable[3] which does not, as required, have zero mean, since the *minimum* effect of tape bending on the measurand is zero.

**Example 5** *Straight-line calibration (re-visited)*

For the straight-line calibration example of Section 4.1 (Example 3), the GUM model constitutes a formula or prescription (not in general necessarily explicit in form) derived from the results of the statistical modelling process. Specifically, the measurement results $\mathbf{a} = (a_1, a_2)^{\mathrm{T}}$ are given in terms of the input quantities $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ by an equation of the form

$$H\mathbf{a} = \mathbf{q}. \tag{4.3}$$

(Compare [22], [1, Clause H.3].) Here, $H$ is a $2 \times 2$ matrix that depends on the values of the $t_i$, and $\mathbf{q}$ a $2 \times 1$ vector that depends on the values of the $t_i$ and the $x_i$.

By expressing this equation as the formula

$$\mathbf{a} = H^{-1}\mathbf{q}, \tag{4.4}$$

a *GUM model* for the parameters of the calibration line is obtained. It is, at least superficially, an explicit expression[4] for the measurement result $\mathbf{a}$. The form (4.3) is also a GUM model, with the measurement result defined *implicitly* by the equation (Section 6).

### 4.2.1 Correlated model parameters

In a range of circumstances some choice is possible regarding the manner in which the input quantities to the model is provided. A group of input quantities can be correlated in that each depends on a common effect. It may be possible to re-express such input quantities so that the common effect appears explicitly as a further input quantity. By doing so, this cause of correlation is *eliminated*, with the potential for a simplification of the analysis. See GUM Clause F.1.2.4. Also, an example in mass comparison [2] illustrates the principle.

[3] This degree of sophistication would not be warranted when measuring the length of a piece of string. It can be important in other applications.

[4] The expression is termed superficially explicit, since the determination of $\mathbf{a}$ *via* a formal matrix inversion is *not* recommended [22]. The form (4.4), or forms like it in other such applications, should not be regarded as an *implementable* formula [22]. Rather, numerically stable matrix factorization algorithms [35] should be employed. This point is not purely academic. The instabilities introduced by inferior numerical solution algorithms can themselves be an appreciable source of uncertainty. It is not straightforward to quantify this effect.

An example of this approach, in the context of measuring the sides of a right-angled triangle, is given in Section 9.4.

In general, the use of modelling principles, before uncertainties are assigned, is often helpful in understanding correlation effects.

### 4.2.2 Constrained uncertainty evaluation

Constrained uncertainty evaluations arise as a consequence of physical limits or conditions associated with the measurands or the model input quantities. Instances include chemical concentrations, departures from perfect form in dimensional metrology and limits of detection.

When chemical concentrations are measured, it will be appropriate to ensure that in cases where all constituent parts are measured the estimates of the measurands sum to unity (or 100%). The associated estimates will inevitably be correlated even if the raw measurements have mutually independent errors.

In assessing the departure from perfect form in dimensional metrology, the measurand is a quantity such as flatness, roundness, perpendicularity, concentricity, etc. These quantities are defined as the *unsigned* departure, assessed in an unambiguously defined way, of a real feature from an ideal feature, and are often very small, but nonzero. Any uncertainty statement associated with such a measurand that is based on a pdf that can embrace zero is physically unrealistic. Such a pdf is invariably asymmetric.

In triangulation, photogrammetry and similar applications, using theodolites, laser interferometers and metric cameras, redundancy of measurement ensures that improved uncertainties are obtained compared with the use of a near-minimal number of measurements. The measurands, point co-ordinates, distances, etc. are interrelated by equality conditions deducible from geometrical considerations.

Within analytical chemistry, measurements of, e.g., trace elements, are made at the limit of detection. At this limit the measurement uncertainty is comparable to the magnitude of the measurement itself. This situation has aspects in common with that in dimensional metrology above, although there are appreciable contextual differences.

The Eurachem Guide to quantifying uncertainty in analytical measurement states

> [32, Appendix F] At low concentrations, an increasing variety of effects becomes important, including, for example,
>
> - the presence of noise or unstable baselines,
> - the contribution of interferences in the (gross) signal
> - ...
>
> Because of such effects, as analyte concentrations drop, the relative uncertainty associated with the result tends to increase, first to a substantial fraction of the result and finally to the point where the (symmetric) uncertainty interval includes zero. This region is typically associated with the practical limit of detection for a given method.
>
> ...

> Ideally, therefore, quantitative measurements should not be made in this region. Nevertheless, so many materials are important at very low levels that it is inevitable that measurements must be made, and results reported, in this region. ... The ISO Guide to the Expression of Uncertainty in Measurement does not give explicit instructions for the estimation of uncertainty when the results are small and the uncertainties large compared to the results. Indeed, the basic form of the 'law of propagation of uncertainties' ... may cease to apply accurately in this region; one assumption on which the calculation is based is that the uncertainty is small relative to the value of the measurand. An additional, if philosophical, difficulty follows from the definition of uncertainty given by the ISO Guide: though negative observations are quite possible, and even common in this region, an implied dispersion including values below zero cannot be "reasonably ascribed to the value of the measurand" when the measurand is a concentration, because concentrations themselves cannot be negative.
>
> ...
>
> Observations are not often constrained by the same fundamental limits that apply to real concentrations. For example, it is perfectly sensible to report an 'observed concentration', that is, an estimate below zero. It is equally sensible to speak of a dispersion of possible observations which extends into the same region. For example, when performing an unbiased measurement on a sample with no analyte present, one *should* see about half of the observations falling below zero. In other words, reports like
>
> observed concentration = $2.4 \pm 8$ mg $l^{-1}$
>
> observed concentration = $-4.2 \pm 8$ mg $l^{-1}$
>
> are not only possible; they should be seen as valid statements.
>
> ...

It is the view of the authors of this guide that these statements by Eurachem are sound. However, this guide takes a further step, related to *modelling* the measurement and through the use of the model defining and making a statement about the measurand, as opposed to the observations. Because a (simple) model is established, this step arguably exhibits even closer consistency with the GUM.

The Eurachem statements stress that observations are not often constrained by the same fundamental limits that apply to real concentrations. It is hence appropriate to demand that the measurand, defined to be the *real* analyte concentration (or its counterpart in other applications) should be constrained to be non-negative. Also, the observations *should not* and *cannot* be constrained, because they are the values actually delivered by the measurement method. Further, again consistent with the Eurachem considerations, assign a pdf to the input quantity, viz., a best (unconstrained) estimate of the analyte concentration, that is symmetric about that value. Thus, the input, $X$, say, is *unconstrained analyte concentration* with a symmetric pdf, and the output, $Y$, say, *real analyte concentration*, with a pdf to be determined.

In terms of these considerations an appropriate GUM model is[5]

$$Y = \max(X,\ 0). \tag{4.5}$$

The rationale behind this simple choice of model is as follows. Should the mean $x$ of the observations prove to be non-negative, it would naturally be taken as the estimate of the measurand $Y$. Such a value would conventionally be used at points removed from the limit of detection. Should $x$ prove to be negative, it cannot be used as a physically feasible estimate of $Y$, since by definition $Y$ is the real analyte concentration and hence non-negative. Taking $y = 0$ in this case is the optimal compromise between the observed values and feasibility (the closest feasible value to the mean observation).

An example illustrating the use of this model is given in Section 9.5.

### 4.2.3 Multi-stage models

Multi-stage models are widespread in metrology. Even the string example (Section 4.2, Example 4) can be interpreted this way. Any situation in which the results and uncertainties from one evaluation become the input quantities and associated uncertainties to a subsequent stage constitute (part of) a multi-stage model. *Within* a model there are frequently sub-models, and therefore multi-staging arises also in this context. Examples abound, especially with calibration.

In the first stage of a multi-stage model, the metrologist is responsible for providing all the input quantities. In subsequent stages, the input quantities constitute some or all of the output quantities from previous stages plus, possibly, further input quantities from the metrologist.

**Example 6** *Example of a multi-stage model, in calibration*

An example of a multi-stage model occurs regularly in calibration, when it is necessary to establish and use a calibration curve. The following description is in the context of Mainstream GUM. There would be an analogous description for circumstances where it was necessary to avoid the Mainstream GUM limitations and perhaps use MCS instead.

Stage 1 involves analysing measurement data that is a function of a second variable, e.g., displacement as a function of applied force. The displacement values, and perhaps the values of the applied force, if they are not known accurately, constitute the (first-stage) model input quantities. The associated uncertainties, and covariances, if relevant, would be assigned. The model specifies the process of fitting a calibration curve to the data to provide the coefficients or parameters of the curve. These parameters constitute the model output quantities. If there is more than parameter (the usual case), their values will almost invariably be correlated, since each parameter will generally be a function of the (same) input data. Thus, the output quantities will have an associated non-diagonal covariance matrix.

Stage 2 involves using the output quantities from Stage 1, viz., the curve parameters and their covariance matrix, as input quantities to a model that constitutes a rule for evaluating the calibration curve for appropriate values of

[5] Related considerations [46, p129] show that if an observation $v$ is $N(\theta,\ 1)$ distributed, i.e., drawn from a Gaussian distribution with mean zero and standard deviation unity, but $\theta \geq 0$, the maximum likelihood estimate of $\theta$ is $\max(v,\ \theta)$.

the argument (force in the above instance).[6] The output quantities will be the values of the curve at these designated points, together with their associated covariance matrix. Again, because these curve values all depend in general on all the inputs, the curve parameters, this covariance matrix will be non-diagonal.

There may be no further stage, since the "interpolated" values provided by the calibration curve may be the primary requirement.

Otherwise, Stage 3 will be the use of the curve values obtained in Stage 2 to provide further measurement results. As an example, take the area under (a specified portion of) the calibration curve. Suppose that this area is to be determined by numerical quadrature because of the impossibility of carrying out the integration analytically. This result can typically be expressed as a linear combination of the curve values provided as input quantities. As another instance, if more than one measurement result is required, e.g., estimates of gradients to the curve at various points, these again can typically be expressed as linear combinations of the curve values. They will, for similar reasons to those above, have a non-diagonal covariance matrix.

The concepts described in Chapter 6 can be applied to the above stages. The various categories within the classification of that chapter would relate to the various types of calibration model, depending on whether it can be expressed explicitly or implicitly or is real-valued or complex-valued. The model is almost always multivariate in the sense of Chapter 6, i.e., it has more than one output. ∎

## 4.3 Assignment of the input probability density functions

The provision of input probability density functions (pdf's) requires the assignment of appropriate statistical distributions (Gaussian, uniform, etc.) to the model input quantities. It can be a challenging step in the formulation stage, Phase 1, of uncertainty evaluation. Valuable guidance is given in the GUM on this matter. Additional aspects are considered here.

Sometimes these input pdf's will be the result of a previous "uncertainty calculation" within the context of a multi-stage model (Section 4.2.3).

In the above straight-line calibration example (Section 4.1, Example 3, and Section 4.2, Example 5), the pdf for each input quantity would have been taken as Gaussian, under the assumption that this knowledge of the errors in the measurement process was available. There would be other types of measurement for which the errors would be expected to be Poissonian, for example. There are many other types of measurement error.

Information concerning the underlying distribution should be deduced in any one instance from all the knowledge that can economically be brought to bear (Section D.2).

[6] In many situations, it is necessary to use the calibration curve *inversely*. Typically, the data in Stage 1 represents a set of *standards*, e.g., established responses to specified concentrations. At Stage 2, it is required to use the calibration curve to determine the concentration corresponding to to an observed response. The mathematical function representing the curve then constitutes an implicit model (Section 6.2.3) (e.g., if the calibration curve is a fifh-degree polynomial).

There is an important class of metrology problems, viz., calibration as above or generally the analysis of experimental data. Suppose that there is a large number of "similar" measurements, such as in the straight-line calibration example in Section 4.1. Suppose also that the errors in these measurements can be taken as uncorrelated. For a calibration function that can be expressed as a linear combination of calibration parameters, as above, these parameters can formally be written as a linear combination of the measured values. For the large number of measurements envisaged, the statistics of the situation are such that almost regardless of the nature of the distribution of the errors in the measurements, a linear combination of the measurements, as here, can be expected to have essentially a Gaussian distribution, as a consequence of the Central Limit Theorem [51, p165]. When there are several such parameters (output quantities) they will almost invariably be correlated, since each is a linear combination of the input quantities. These paramaters would be described by a multivariate (joint) Gaussian distribution. See Section 4.3.2.

The straight line above would have an intercept and a gradient that are correlated.[7]

Even if the calibration function depends nonlinearly on its parameters, by linearizing this function about the solution values of the parameters, to first order similar considerations apply as in the linear case [22]. In cases of doubt the validation procedures of Section 8 should be undertaken to determine whether linearization is justified.

The GUM [1] discriminates between the Type A evaluation of uncertainty—that based on statistical means—and the Type B evaluation of uncertainty—that based on non-statistical means. Although this terminology is sometimes used in this guide for alignment with the GUM, no great distinction is made here, since all types of uncertainties can be classified by appealing to a unifying principle (Section D.2). It is sometimes more useful to examine the distinction between errors in a measurement that can be regarded as random and those that can be regarded as systematic.[8] In some instances a systematic error can be treated as a bias and handled as part of statistical modelling (Sections 4.1 and 9.4).

Recommended assignments for Type A evaluations of uncertainty for three important types of input quantity follow. They are based on the application of the Principle of Maximum Entropy (Section D.2). For each type the assignment applies on the condition that no information other than that specified is available. A "small number" of repeated observations is taken as ten or fewer. A "large number" is taken as greater than 20.

[7] It is possible in some applications such as this one to re-express (re-parametrise) the straight line such that its parameters are uncorrelated [22]. Also see Section 4.2.1. The example in Clause H.3 of the GUM illustrates this point. Such re-parametrisation is not *always* a practical proposition, however, because of the conflict between a numerically or statistically convenient representation and the requirements of the application. However, the possibility of re-parametrization should always be considered carefully for at least two reasons. One reason is that the result corresponding to a sound parametrization can be obtained in a numerically stable manner [22], whereas a poor parametrization can lead to numerically suspect results. Another reason is that a poor parametrization leads to *artificially large* correlations in the output quantities. Decisions about the natural correlation present in the results cannot be made in terms of these *induced* correlation effects.

[8] The subdivision into Type A and Type B evaluations of uncertainty will correspond in some instances to random and systematic effects, respectively, but not in all circumstances.

**A small number of observations.** The small number of repeated observation is regarded as being drawn from an unknown distribution. Take the mean and standard deviation of the mean. Assign a Gaussian pdf with these parameters to the input quantity.

**A small number of observations drawn from a Gaussian distribution.** The small number of repeated observation is regarded as being drawn from a Gaussian distribution. Take the mean and standard deviation of the mean. Assign a Student's-$t$ pdf with these parameters, with a number of degrees of freedom one less than the number of observations, to the input quantity.

**A large number of observations.** A different attitude can legitimately be taken if a set of repeated observations is reasonably large. The observations, when portrayed as a histogram, for instance, might indicate features such as asymmetry or bimodality. If no additional information is available, rather than using the PME, the data itself can be used to "define" its own pdf. Specifically, the pdf is given by assigning a probability $1/q$, where $q$ is the number of observations, to each value of the input quantity that is equal to one of these observations, and zero probability to all other values of the input quantity. Sampling from the pdf would take place simply by selecting a data value at random, with equal probability attached to all values. Any value selected is replaced before a further value is sampled.

**An intermediate number of observations.** The treatment of an intermediate number of observations is more problematical. It will be adressed in a future edition of this guide.

### 4.3.1 Univariate Gaussian distribution

There are many circumstances where measurements contain errors that stem from a large number of sources and no one contribution dominates. In these situations it is reasonable to regard the measurements as having Gaussian errors. One common instance is a parameter arising from least-squares fitting (Section 4.3).

The (univariate) *Gaussian* or *normal* distribution with mean $\mu$ and standard deviation $\sigma$ in the variable $X$ has the pdf

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-(x-\mu)^2/(2\sigma^2)\right\}, \qquad -\infty < x < \infty.$$

The *standardized Gaussian* distribution in the variable $Z$, with zero mean and unit standard deviation, is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \qquad -\infty < z < \infty.$$

Its distribution function, denoted by $\Phi(z)$, is

$$\Phi(z) = \int_{-\infty}^{z} \phi(v)dv.$$

The probability that $X$ lies between $c$ and $d$, where $c < d$, is

$$\begin{aligned}\frac{1}{\sigma\sqrt{2\pi}}\int_c^d \exp\left\{-(x-\mu)^2/(2\sigma^2)\right\}dx &= \int_{(c-\mu)/\sigma}^{(d-\mu)/\sigma} \exp(-z^2/2)\,dz \\ &= \Phi((d-\mu)/\sigma) - \Phi((c-\mu)/\sigma).\end{aligned}$$

The inverse function $\Phi^{-1}(p)$ gives the value of $z$ such that $\Phi(z) = p$, a stated probability.

Tables and software for $\Phi$ and its inverse are widely available.

### 4.3.2 Multivariate Gaussian distribution

In general, multivariate distributions can be defined in terms of joint probability density functions $g(\mathbf{x})$. The *multivariate Gaussian* distribution (or multinormal distribution) with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ and covariance matrix $V$ of order $p$ has pdf

$$g(\mathbf{x}) = \frac{1}{(\det(2\pi V))^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} V^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

The *set* of parameters arising from least-squares fitting (Section 4.3.1) can often be described by such a distribution.

### 4.3.3 Univariate uniform distribution

It is often assumed that when a model input quantity is given in a manufacturer's specification in the form of a "plus/minus accuracy statement", the corresponding pdf should be taken as uniform with limits dictated by the accuracy statement. If there is no other information available, this attitude is consistent with the Principle of Maximum Entropy (PME) (Appendix D).

The *uniform* or *rectangular* distribution has pdf

$$g(x) = \begin{cases} 1/(b-a), & a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

It states that any value of $x$ in the interval $[a,\ b]$ is equally probable and that the probability of a value of $x$ outside this interval is zero.

Consider two values $c$ and $d$, where $c < d$. The probability that $X$ lies between $c$ and $d$ is straightforwardly confirmed to be

$$\int_c^d g(x)dx = \begin{cases} 0, & d \le a, \\ (d-a)/(b-a), & c \le a \le d \le b, \\ (d-c)/(b-a), & a \le c < d \le b, \\ (b-c)/(b-a), & a \le c \le b \le d, \\ 0, & b \le c. \end{cases}$$

The authors of this guide regard the above assumption with some scepticism. Although the PME is perfectly reasonable in the light of such information, it is the information that leads to its being invoked in this manner that should be questioned. The belief in uniform distributions as being "right" would appear to be an overused and misused belief. It can lead to anomalies.

Can taking a uniform pdf be a better model in general than using, say, a Gaussian? There are indeed genuine instances for the use of a uniform pdf. An

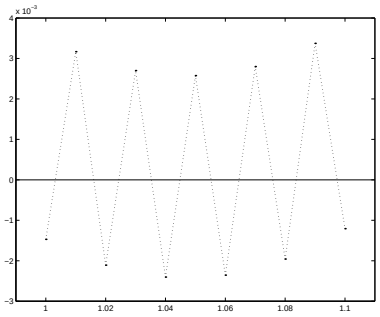


Figure 4.1: The errors in successive values displayed by a simulated instrument having a resolution of two decimal places. The values shown are the differences between the values of $\sin t$ that would be displayed by the instrument and the actual values of $\sin t$, for $t = 1.00, 1.01, \ldots, 1.10$ radians.

example is the digital resolution of an instrument, in which the error can be regarded as being equally likely anywhere within plus or minus half a unit in the last displayed digit.[9]

The quantization error in analogue to digital conversion also falls (with some exceptions) into this category. There would appear to be few other genuine examples. It would be desirable, especially in a competitive environment or when particularly reliable uncertainty statements are required, to approach suppliers to relate the provided accuracy statement to the context in which it was made. The supplier might, for example, be speaking loosely, e.g., to imply a 99% coverage interval, say, with the previously unmentioned information that an underlying Gaussian pdf was reasonable. The contextual information might relate, for example, to reject rates in a production process.

Information is available [10] on a method for reducing the uncertainty associated with instrument resolution when a series of observations is taken. It involves randomizing the zero setting, where this is possible, before taking each observation. The mean of a set of $q$ observed values so obtained can be expected to have an uncertainty that is smaller than that of an individual observation by a factor of $\sqrt{q}$. This result is to be compared with conventional repeated observations in situations where the uncertainties are dominated by those of the instrument resolution: the mean of the observations has no better property than the individual observations.

[9] This statement is correct for a single reading. There are additional considerations for a sequence of readings corresponding to a slowly varying signal. The errors in the resolved sequence are *serially correlated* as a consequence of the resolution of the instrument. Figure 4.1 shows the errors in successive values displayed by a simulated instrument having a resolution of two decimal places. The values shown are the differences between the values of $\sin t$ that would be displayed by the instrument and the actual values of $\sin t$, for $t = 1.00, 1.01, \ldots, 1.10$ radians. Any analysis of such data that did not take account of the very obvious serial correlation would yield a flawed result. The effects of serial correlation depend on the relative sizes of the uncertainties in the signal, the instrument resolution and the magnitudes of the changes in successive values of the signal (the last-mentioned item depending on the sampling rate). In hopefully many cases they will be negligible, but it is appropriate to establish when this is indeed the case. In the context of calibration it is stated [28], but the point is more general, that the measurement uncertainty associated with the calibration of all low-resolution indicating instruments is dominated by the finite resolution provided this resolution is the only dominant source in the uncertainty budget.

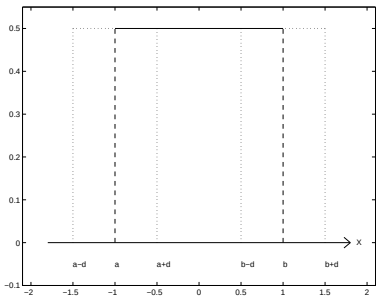


Figure 4.2: A uniform pdf with inexact endpoints. The diagram is conceptual: the "height" of the pdf would in fact vary with the endpoints in order to maintain unit area.

### 4.3.4 Inexactly specified uniform distributions

Consider a random variable $X$ having nominally a uniform pdf, specified in terms of its lower limit $a$ and upper limit $b$. These endpoints may be known inexactly. For instance, suppose the values $a = -1$ and $b = 1$ are given, only the quoted figures are reliable, and no other information is available. Then, it can be concluded that the actual value of $a$ lies between $-1.5$ and $-0.5$ and $b$ between 0.5 and 1.5.[10] Thus, $X$ in fact lies in the broader interval $[-1.5,\ 1.5]$ rather than $[-1,\ 1]$. See Figure 4.2. How important is this consideration in practice? In what manner is $X$ distributed over this interval?

These considerations are a direct counterpart of those in the GUM in which an input standard uncertainty is obtained from a Type B evaluation and cannot be treated as exactly known. See GUM Clause G.4.2. There the inexactness is manifested as a finite number of effective degrees of freedom.

Suppose that the left endpoint is regarded as lying in the interval $[a-d,\ a+d]$ and the right endpoint in $[b-d,\ b+d]$. It is assumed that "the $d$" is the same for each endpoint. The treatment can be generalised if needed. It is henceforth assumed that the actual value of the left endpoint is equally likely to lie *anywhere* in $[a-d,\ a+d]$, with a similar statement for the right endpoint.[11] Thus, the left and right endpoints are taken as uniform random variables, $A$ and $B$, say. It follows that

$$X = A + (B - A)V,$$

where $A$ is uniform over $[a-d,\ a+d]$, $B$ is uniform over $[b-d,\ b+d]$ and $V$ is uniform over $[0,\ 1]$.

A Monte Carlo Simulation using the introductory example, viz., with $a = -1.0$, $b = 1.0$ and $d = 0.5$ gave the histogram in Figure 4.3, as a scaled estimate of the pdf of $X$.

Note the "shape" of the pdf. It is uniform over the region between the inner extremities of the inexact endpoints, i.e., where there is no doubt concerning

[10] If instead the values $a = -1.0$ and $b = 1.0$ were quoted, it would be concluded that the left endpoint were between $-1.05$ and $-0.95$ and the right endpoint between 0.95 and 1.05.

[11] An alternative approach can be used. It could be assumed, for instance, that each endpoint can be regarded as a Gaussian (rather than a uniform) variable, centred on that endpoint, with a stated standard deviation. The analysis and the result would differ from that here. The choice of approach would be made using expert judgement.

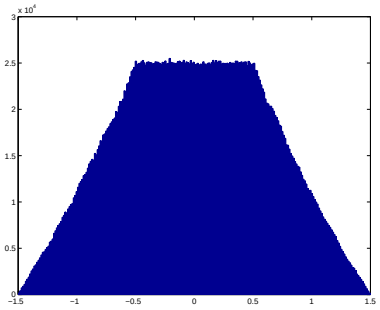

Figure 4.3: A histogram produced using Monte Carlo Simulation for the model $X = A + (B - A)V$, where $A$ is uniform over $[a - d,\ a + d]$, $B$ is uniform over $[b - d,\ b + d]$ and $V$ is uniform over $[0,\ 1]$, with $a = -1.0$, $b = 1.0$ and $d = 0.5$. Figure 4.2 refers. The histogram provides a scaled estimate of the pdf of $X$. It corresponds to an input quantity that is defined by a uniform distribution between inexact limits, theirselves being represented by uniform distributions.

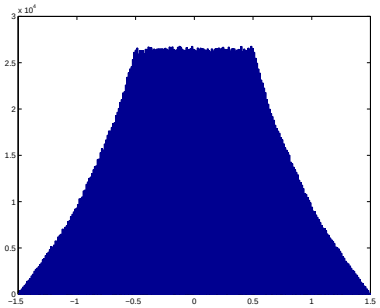

Figure 4.4: As Figure 4.3 except that the endpoints are related as described in the text.

their location. Between the inner and outer extremities it reduces from the uniform height to zero in what is approximately a quadratic manner. Beyond the outer extremities the pdf is zero. The piecewise nature of the pdf is comparable to that for the sum of uniform pdf's, where the pieces form polynomial segments [26].

The standard deviation of $X$, assuming the exactness of the endpoints (equivalent to taking $d = 0$), is $1/\sqrt{3} = 0.577$. That for the above finite value of $d$ is 0.625. As might be expected, the inexactness of the endpoints increases the value. The extent to which this increase (8%) is important depends on circumstances.

There will be situations where the inexact endpoints would be expected to "move together", i.e., the knowledge of one of them would imply the other. In this circumstance the pdf of $X$ is slightly different. See Figure 4.4. The standard deviation of $X$ is now 0.600 (a 4% increase over that for $d = 0$), roughly halfway between that for the above pdf and the pure uniform pdf. The flanks of the pdf now have greater curvature.

The final statement to be made here concerning uniform distributions with inexactly defined endpoints is that the effects of such endpoints on the evaluation

of uncertainty increase with the relative amount of inexactness. This point is qualitatively consistent with the use of an effective number of degrees of freedom, as above, in the GUM. Increased inexactness will give rise to a smaller number and yield greater uncertainty through a larger coverage factor from the Student's-$t$ distribution.

The main message is that inexactness in the information that leads to assigning pdf's not only modifies the forms of those pdf's, but influences the relevant standard deviations.

### 4.3.5 Taking account of the available information

It is beyond the scope, in this first edition of this best-practice guide, to state how all information available can properly be taken into account. Some remarks are made, however, indicating how the Principle of Maximum Entropy can be used to advantage. A future edition will provide more details.

If a set of values, in the form of repeated observations, is available, and *no other information* is provided, the PME would mandate the use of a Gaussian pdf with mean and standard deviation deduced from the measurements.

If only a lower and an upper limit were available the PME would support the choice of a uniform pdf, as above.

Suppose a prior pdf, say a Gaussian, were available, perhaps from historical information such as that obtained in previous calibrations. Suppose that further measurements were available. The use of the PME would permit both sources of information to be combined to deliver a Students-$t$ pdf that could be expected to be more reliable than the pdf from either source alone.

Suppose a prior uniform pdf were available, perhaps from sound knowledge of limits, and that one measurement was made. From the PME, the uniform pdf that would be inferred from the limits alone would be modified by the measurement. The GUM provides (in GUM Clause 4.3.8) the pdf in this case.

Other cases can be handled, and give superior results in general than if treated without taking account of the available information. Such cases may require the services of a statistician.

Appendix D considers some of the issues involved.

It is relevant to note that in the context of Mainstream GUM, which works only with the standard deviations (and the means) of the input pdf's, the GUM states

> [GUM Clause E.4.2] When the standard uncertainty of an input quantity cannot be evaluated by analysis of the results of an adequate number of repeated observations, a probability distribution must be adopted based on knowledge that is much less extensive than might be desirable. That does not, however, make the distribution invalid or unreal; like all probability distributions it is an expression of what knowledge exists.

This attitude is consistent with a Bayesian view [65].

## 4.4 Determining the output probability density function

The pdf of the output quantity is completely defined by the model together with the assigned input pdf's. Appropriate analysis or calculation is needed, however, to determine it. Chapter 5 covers candidate approaches for forming the output pdf in the univariate case, and indicates its counterpart in the multivariate case.

## 4.5 Providing a coverage interval

The provision of a coverage interval is the use of the pdf of the output quantity to determine a lower limit and an upper limit of an interval that can be expected to contain 95% (or some other specified proportion) of the values that can be reasonably be attributed to the measurand. See Chapter 5 for methods for determining the pdf of the output quantity. See Section 2.3 for information on coverage intervals. Coverage intervals can be obtained objectively from a pdf. They can also be obtained from *coverage factors* and an *assumption* concerning the pdf.

### 4.5.1 Coverage intervals from distribution functions

If the distribution function is known, a coverage interval can be obtained as indicated in Section 2.3.

### 4.5.2 Coverage intervals from coverage factors and an assumed form for the distribution function

The Mainstream GUM approach (see GUM Clause G.1.1) to determining a coverage interval is as follows. The aim (using the notation of the GUM) is to provide, using the estimate $y$ of the measurand $Y$ and the standard uncertainty $u(y)$ of the estimate, an expanded uncertainty $U_p = k_p u(y)$. With the estimate $y$, this value $U_p$ defines an interval $[y - U_p,\ y + U_p]$ that has a specified coverage probability $p$.

In summarizing its recommendations for determining this coverage interval, the GUM states:

> [GUM Clause G.6.1] The coverage factor $k_p$ that provides an interval having a level of confidence $p$ close to a specified level can only be found if there is extensive knowledge of the probability distribution of each input quantity and if these distributions are combined to obtain the distribution of the output quantity. The input estimates $x_i$ and and their standard uncertainties $u(x_i)$ by themselves are inadequate for this purpose.

Further,

> [GUM Clause G.6.2] Because the extensive computations required to combine probability distributions are seldom justified by the extent and reliability of the available information, an approximation to the distribution of the output quantity is acceptable. Because of the

> Central Limit Theorem, it is usually sufficient to assume that the probability distribution of $(y - Y)/u_C(y)$ is the $t$-distribution and take $k_p = t_p(\nu_{\text{eff}})$, with the $t$-factor based on an effective degrees of freedom $\nu_{\text{eff}}$ of $u_C(y)$ obtained from the Welch-Satterthwaite formula ...

The statement[12] concerning the extensive computation to combine probability distributions is no longer tenable, with PCs faster than 1 GHz being commonplace. Unless the model is complicated, the determination of the pdf of the output quantity and hence the required coverage interval to the required number of figures, can, with todays' PCs, be carried out in computation times of minutes or even seconds (Section 7.3.1).

### 4.5.3 An approximation is acceptable, but is it an acceptable approximation?

The statement from the GUM reproduced in Section 4.5.2 concerning the Central Limit Theorem demands investigation. It is accepted that it is *usually* sufficient to assume that the output is distributed as Student's-$t$. The difficulty lies in deciding when this assumption can be made. The GUM offers no *specific* guidance in this regard. This document supports that approach when it can be justified, but recommends that in any case of doubt the validation approach of Section 8.2 should be employed.

However, the GUM does provide some advice regarding the circumstances when Mainstream GUM can be expected to hold:

> [GUM Clause 6.6] For many practical measurements in a broad range of fields, the following conditions prevail:
>
> - the estimate $y$ of the measurand $Y$ is obtained from estimates $x_i$ of a significant number of input quantities $X_i$ that are describable by well-behaved probability distributions, such as the normal and rectangular distributions;
> - the standard uncertainties $u(x_i)$ of these estimates, which may be obtained from either Type A or Type B evaluations, contribute comparable amounts to the combined standard uncertainty $u_C(y)$ of the measurement result $y$;
> - the linear approximation implied by the law of propagation of uncertainty is adequate (see 5.1.2 and E.3.1);
> - the uncertainty of $u_C(y)$ is reasonably small because its effective degrees of freedom $\nu_{\text{eff}}$ has a significant magnitude, say greater than 10.
>
> Under these circumstances, the probability distribution characterized by the measurement result and its combined standard uncertainty can be assumed to be normal because of the Central Limit Theorem; and $u_C(y)$ can be taken as a reasonably reliable estimate of the standard deviation of the normal distribution because of the significant size of $\nu_{\text{eff}}$.

[12] The GUM uses the notation $u_C(y)$ for *combined standard uncertainty*, i.e., that associated with $y$. This guide simply uses $u(y)$.

This advice is sound in a qualitative sense but, again, it is unclear when the circumstances hold. The problem is that the distinction between Phase 1 and Phase 2 of uncertainty evaluation[13], as indicated in Section 4, becomes blurred. The intention of the subdivision into the two phases is to permit all decisions to be made in Phase 1 and the mechanical calculations to be made in Phase 2.

In terms of the set of conditions in GUM Clause 6.6, listed above, it is unclear what is meant by

- "a significant number of input quantities"
- "well-behaved probability distributions"
- the standard uncertainties of the $x_i$ contributing comparable amounts[14]
- the adequacy of linear approximation, and
- the output uncertainty being reasonably small.

The concern is that because none of these considerations is explicitly quantified, different practitioners might adopt different interpretations of the same situation, thus causing divergence of results.

Further, an approximation is acceptable, but is it an acceptable approximation?

## 4.6 When the worst comes to the worst

Consider a situation in which no assumption is to be made about the pdf of the output quantity other than an estimate of its mean value $y$ and the standard deviation $u(y)$ of this estimate. One reason for wishing to make no assumption is that it may be difficult or impossible to obtain distributional information about some of the input quantities and it is deemed inappropriate to invoke the Principle of Maximum Entropy. In such a circumstance, a *conservative* estimate of a coverage interval can be obtained using some traditional results from the statistical literature.[15] Two results are possible. One result is general, applying to all distributions. The other relates to instances in which one is prepared to make a single assumption, viz., that the distribution is symmetric.

### 4.6.1 General distributions

Suppose that it is required to quote an uncertainty interval for the measurand $Y$ corresponding to a coverage probability of 95%, and that *nothing* is known about the distribution.

[13]The JCGM/WG1 has agreed that this distinction is useful. It will form part of a supplemental guide being produced by JCGM/WG1.

[14]This statement is taken here to mean that the standard uncertainties of the input quantities, when scaled by the magnitudes of the corresponding sensitivity coefficients, contribute comparable amounts.

[15]Such an estimate is inconsistent with the intention of the GUM which promotes the use of a *realistic* coverage interval:

> [GUM, Clause 0.4] . . . the ideal method for evaluating and expressing uncertainty in measurement should be capable of readily providing such an interval, in particular, one with a coverage probability or level of probability that corresponds in a realistic way with that required.

There may, however, be special situations where a conservative estimate is useful.

The coverage interval $y \pm ku(y)$, where $k = 4.47$, contains at least 95% of the distribution of $y$-values.

This result is derived from *Chebyshev's inequality* which states that the probability that $Y$ lies in the interval $y \pm ku(y)$ is at least $1 - k^{-2}$. The value of $k$ for which $1 - k^{-2} = 0.95$ is 4.47. It is stressed that this result applies *regardless of the distribution.* By its nature it cannot be as sharp as an interval derived from knowledge of the pdf of $y$, e.g.,

- If $Y$ is uniform this interval is $y \pm 1.65u(y)$
- If $Y$ is Gaussian it is $y \pm 1.96u(y)$.

The length of the interval derived from Chebyshev's inequality is 2.7 times the length of that for uniform $Y$ and 2.3 times that for Gaussian $Y$.

*Note.* These results applies only if the number of degrees of freedom is infinite, or in practice large. Otherwise, the $k$-factor becomes inflated, as in the case of Student's-$t$ distribution [53].

### 4.6.2 Symmetric distributions

If it is known that the distribution is *symmetric*, tighter results based on *Gauss's inequality* are possible.

The coverage interval $y \pm ku(y)$, where $k = 2.98$, contains at least 95% of the distribution of $y$-values.

Gauss's inequality states that the probability that $Y$ lies in the interval $y \pm ku(y)$ is at least $1 - \frac{4}{9}k^{-2}$. The value of $k$ for which $1 - \frac{4}{9}k^{-2} = 0.95$ is 2.98.

It is noted that this interval is only approximately 50% longer than that when $Y$ is Gaussian (Section 4.6.1).

*Note.* These results applies only if the number of degrees of freedom is infinite, or in practice large. Otherwise, the $k$-factor becomes inflated, as in the case of Student's-$t$ distribution.

### 4.6.3 Making stronger assumptions

Tighter, distribution-free, coverage intervals can be obtained if stronger assumptions are made, such as unimodality (single-peakedness) as well as symmetry of the pdf.

# Chapter 5

# Candidate solution approaches

This chapter covers candidate solution processes for the calculation phase, Phase 2, of the uncertainty evaluation problem formulated in Section 3.1. The starting point is (i) the availability of a model $f$ or $\mathbf{f}$ that relates the input quantities $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ to the scalar measurand $Y$ or vector measurand $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\mathrm{T}}$ through $Y = f(\mathbf{X})$ or $\mathbf{Y} = \mathbf{f}(\mathbf{X})$, and (ii) assigned probability density functions (pdf's) $g_1(X_1), \ldots, g_n(X_n)$ for the input quantities. If the input quantities are correlated, they will have a joint pdf.

It is required to determine the pdf $g(Y)$ for the output quantity $y$ or the (joint) pdf $\mathbf{g}(\mathbf{Y})$ for $\mathbf{Y}$.

Once $g(Y)$ has been obtained a 95% coverage *interval* for the (scalar) measurand $Y$ can be derived. Once $\mathbf{g}(\mathbf{Y})$ has been obtained, a 95% coverage *region* for the (vector) measurand $\mathbf{Y}$ can be derived.

Three approaches to the determination of the output pdf for $Y$ or $\mathbf{Y}$ are considered and contrasted:

1. Analytical methods
2. Mainstream GUM
3. Numerical methods.

All three approaches are consistent with the GUM. Mainstream GUM is the procedure that is widely used and summarized in GUM Clause 8. Analytical methods and numerical methods fall in the category of "other analytical and numerical methods" (GUM Clause G.1.5). Under the heading of "Analytical methods" below, mention is also made of "Approximate analytical methods".

## 5.1 Analytical methods

*Analytical* methods to obtain the pdf of $Y$ or $\mathbf{Y}$ are preferable in that they do not introduce any approximation, but can be applied in relatively simple cases only. Their application may require the services of an expert. A treatment of such methods, based essentially on the use of formula (3.1) is available [26]. Instances

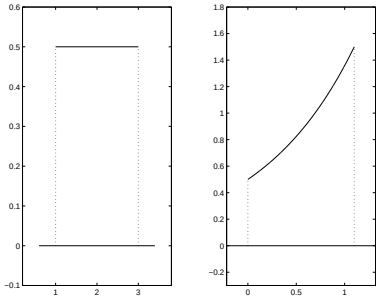

Figure 5.1: A uniform probability density function (left) for the input $X$ and the corresponding probability density function for the output $Y$, where $Y$ is related to $X$ by the model $Y = \log(X)$.

that can be so handled include additive models, $Y = a_1X_1 + \ldots + a_nX_n$, where all $X_i$ are Gaussian or all are uniform. In the latter case, unless $n$ is small, the multipliers $a_i$ must be equal and the semi-widths of the uniform pdf's identical[1] to avoid formidable algebra.

### 5.1.1 Single input quantity

The case of a single input quantity ($n = 1$) is amenable to analytic treatment [51, pp57-61]. If the model function $f(X)$ is differentiable and strictly monotonic, $Y$ has the pdf

$$g(y) = g(f^{-1}(y))|d(f^{-1}(y))/dy|. \tag{5.1}$$

**Example 7** *A logarithmic transformation*

If the model is $Y = \ln(X)$ with $X$ having uniform pdf in $[a,\ b]$, the application of Formula (5.1) gives

$$G(y) = \begin{cases} 0, & y \le \ln(a), \\ (\exp(y) - a)/(b - a), & \ln(a) \le y \le \ln(b), \\ 1, & \ln(b) \le y. \end{cases}$$

(cf. Section 4.3.5). Figure 5.1 depicts the uniform pdf (left) for $X$ and the corresponding pdf for $Y$ in the case $a = 1$, $b = 3$.

This case is important in, say, electromagnetic compatibility measurement, where conversions are often carried out between quantities expressed in linear and decibel units using exponential or logarithmic transformations [62]. ∎

**Example 8** *A linear combination of Gaussian distributions*

Suppose the model is

$$Y = a_1X_1 + \cdots + a_nX_n,$$

where $a_1, \ldots, a_n$ are specified constants, and, for $i = 1, \ldots, n$, $X_i$ has a Gaussian distribution with mean $\mu_i$ and standard deviation $\sigma_i$. Then $Y$ has a Gaussian

[1] In this case $Y$ is a B-spline with uniform knots [18].

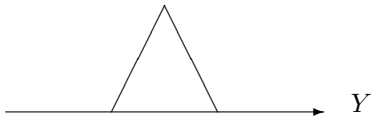


Figure 5.2: The probability density function for the sum $Y = X_1 + X_2$ of two uniform distributions $X_1$ and $X_2$ of identical semi-widths.

distribution with mean $\mu = a_1\mu_1 + \cdots + a_n\mu_n$ and standard deviation $(a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2)^{1/2}$.

**Example 9** *The sum of two uniform distributions with the same semi-widths*

Suppose the model is

$$Y = X_1 + X_2$$

and, for $i = 1, 2$, $X_i$ has a uniform pdf with mean $\mu_i$ and semi-width $a$ (and hence standard deviation $a/\sqrt{3}$). Then $Y$ has a symmetric triangular pdf $g(y)$ with mean $\mu = \mu_1 + \mu_2$, semi-width $2a$ and standard deviation $a\sqrt{2/3}$. Geometrically, this pdf, for the case $\mu_1 + \mu_2 = 0$, takes the form

$$g(y) = \begin{cases} 0, & y \le -2a, \\ (2a + y)/(4a^2), & -2a \le y \le 0, \\ (2a - y)/(4a^2), & 0 \le y \le 2a, \\ 0, & 2a \le y. \end{cases}$$

For general $\mu_1$ and $\mu_2$, the pdf is the same, but centred on $\mu_1 + \mu_2$ rather than zero. Geometrically, this pdf takes the form indicated in Figure 5.2.

**Example 10** *The sum of n uniform distributions of the same semi-width*

Suppose the model is

$$Y = X_1 + \cdots + X_n$$

and, for $i = 1, \ldots, n$, $X_i$ has a uniform pdf with mean $\mu_i$ and semi-width $a$ (and hence standard deviation $a/\sqrt{3}$). Then $Y$ is a B-spline of order $n$ (degree $n-1$) with mean $\mu = \mu_1 + \cdots + \mu_n$ and standard deviation $a\sqrt{n/3}$.

**Example 11** *The sum of two uniform distributions of arbitrary semi-widths*

Suppose the model is

$$Y = \alpha_1 X_1 + \alpha_2 X_2$$

and, for $i = 1, 2$, $X_i$ has a uniform pdf with mean $\mu_i$ and semi-width $a_i$ (and hence standard deviation $a_i/\sqrt{3}$). Then $Y$ has a symmetric trapezoidal pdf $g(y)$ with mean $\mu = \alpha_1\mu_1 + \alpha_2\mu_2$, semi-width $\alpha_1 a_1 + \alpha_2 a_2$ and standard deviation $\{(\alpha_1^2 a_1^2 + \alpha_2^2 a_2^2)/3\}^{1/2}$. Geometrically, this pdf takes the form indicated in Figure 5.3.

Analytical solutions in some other simple cases are available [26, 28].


Figure 5.3: The probability density function for a general linear combination $Y = \alpha_1 X_1 + \alpha_2 X_2$ of two uniform distributions $X_1$ and $X_2$. It is symmetric about its midpoint.

### 5.1.2 Approximate analytical methods

Approximate analytical methods are approaches that fall part-way between the analytical methods of Section 5.1 and the numerical methods of Section 5.3. They are related to Mainstream GUM (Section 6), but take the analysis further in order to provide approximate analytical expressions for the pdf of the output quantity in cases where Gaussian or Student's-$t$ pdf's obtained in the conventional way would be invalid.

A treatment [28] of some calibration examples using approximate analytic methods provides output pdf's in the form of

1. A *uniform output pdf* in the calibration of a hand-held digital multimeter,
2. A *symmetric trapezoidal pdf* in the calibration of a vernier caliper,
3. A further *trapezoidal pdf* in the calibration of a temperature block calibrator.

The first of these examples is used subsequently in this guide (Section 9.3) in the context of the MCS approach to uncertainty evaluation and the results compared with those of [28] and Mainstream GUM.

## 5.2 Mainstream GUM

The GUM makes the following statement about the pdf's of the input quantities:[2]

> [GUM Clause 4.1.6] Each input estimate $x_i$ and its associated uncertainty $u(x_i)$ are obtained from a distribution of possible values of the input quantity $X_i$. This probability distribution may be frequency based, that is, based on a series of observations $x_{i,k}$ of $X_i$, or it may be an *a priori* distribution. Type A evaluations of standard uncertainty components are founded on frequency distributions while Type B evaluations are founded on *a priori* distributions. It must be recognized that in both cases the distributions are models that are used to represent the state of our knowledge.

[2] To this statement, the comment must be added that some pdf's may be based on *both* types of information, viz., prior knowledge *and* repeated observations. Evaluations of standard uncertainty in this setting are not purely Type A or Type B. The GUM gives one such instance (GUM Clause 4.3.8, Note 2.)

The intent of Mainstream GUM is to derive the parameters of a Gaussian or Student's-$t$ distribution that represents the output pdf $g(y)$ given the pdf's of the inputs $X_i$ to the model.

For Mainstream GUM, the following steps constitute the calculation phase, Phase 2:

1. Obtain from the (joint) input pdf(s) the means $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ and the standard deviations $\mathbf{u}(\mathbf{x}) = (u(x_1), \ldots, u(x_n))^{\mathrm{T}}$ of the input quantities.

2. Form the covariances $u(x_i, x_j)$ of the input quantities from the appropriate joint pdf's for any pair of correlated input quantities.

3. Form the partial derivatives of first order of the model with respect to the input quantities. See Section 5.5.

4. Form numerically the (mean) measurement result by evaluating the model at the estimates of the input quantities (the means of the input pdf's).

5. Form the model sensitivities as the values of the partial derivatives evaluated at these means. See Section 5.5.

6. Determine the combined standard uncertainty, viz., the standard deviation of the model output quantity, by combining the values of the input standard uncertainties, the input covariances and the model sensitivities (GUM Formula (13)). See Section 6.2.1.

7. Calculate $\nu$, the effective number of degrees of freedom of the output quantity, using the Welch-Satterthwaite formula (GUM Equation (G.2b)).[3]

8. Compute the expanded uncertainty, and hence an interval having a stipulated coverage probability containing the measurand, by forming the appropriate multiple of the combined standard uncertainty under the assumption that the output is a Gaussian distribution ($\nu = \infty$) or a Student's-$t$ distribution ($\nu < \infty$).

A review of the Mainstream GUM procedure is given in Section 5.4. Details, procedures and examples are given in Chapter 6.

## 5.3 Numerical methods

It would rarely be a practical proposition to use the integral expression (3.1) in Section 3.1 as the basis for the numerical determination of $g(y)$, the pdf of the output quantity. A multivariate quadrature rule[4] would need to be devised that was capable of delivering $g(y)$ to a prescribed numerical accuracy for each choice of $y$. Further, the quadrature rule would have to be applied at a sufficiently fine set of $y$-values to provide $g(y)$ adequately.

[3] Mainstream GUM does not state how $\nu$ is to be calculated when the input quantities are correlated.

[4] A quadrature rule is a numerical integration procedure. Examples in the univariate case are the trapezoidal rule and Simpson's rule.

### 5.3.1 Monte Carlo Simulation

Rather than attempting to evaluate the integral (3.1), Monte Carlo Simulation (MCS) [4, 19, 21, 23, 54, 64] encompasses an entirely different approach, based on the following considerations. The expected value of the output quantity $Y$ is conventionally obtained by evaluating the model for the estimated (mean) values $x_1, \ldots, x_n$ of the input quantities to give the value $y$. However, since each input quantity is described by a pdf rather than a single number, a value as legitimate as its mean can be obtained by drawing a value at random from this function.

MCS operates in the following manner,[5] based on this consideration. Generate a value at random from the pdf for each input quantity and form the corresponding value of the output, by evaluating the model for these values as input quantities. Repeat this process many times, to obtain in all $M$, say, estimates of the output quantity. According to the Central Limit Theorem [51, p169], the mean value $y$ of the estimates of the output quantity obtained in this manner converges as $1/M^{1/2}$, if the standard deviation $u(y)$ of $y$ exists. Irrespective of the dimensionality of the problem, i.e., the number $n$ of input quantities, it is (only) necessary to quadruple $M$ in order to halve the expected uncertainty in the estimate of $u(y)$. Standard numerical quadrature would require a factor of $2^{N/2}$ for this purpose. Thus, the basic concept of MCS has reasonable convergence properties. It is straightforward to implement for simple or even moderately complicated problems. Its *general* implementation requires effort: see Section 7.7. A broad introduction to MCS is available [38], as is a discussion on uncertainty propagation in Monte Carlo calculations [57].

Details, procedures and examples are given in Chapter 7.

## 5.4 Discussion of approaches

The approach used for any particular problem needs to be chosen with care. As indicated, Mainstream GUM is the "method of choice" for many organizations. Analytical methods are in a sense ideal when applicable. Numerical methods offer flexibility. MCS is increasingly used by laboratories and industrial organizations. A more detailed comparison of the three approaches will be prepared as part of the SS*f*M programme (Section 1.1) and given in a future edition of this guide. Here, selected comments are made.

### 5.4.1 Conditions for Mainstream GUM

Mainstream GUM requires

1. The linearization of the model to be sufficiently accurate.[6]

2. The applicability of the Central Limit Theorem (GUM Clause G.2.1), implying the representativeness of $g(y)$, the pdf of the output quantity,

[5] This description applies to a model with a single output quantity. For a multivariate problem, additional considerations apply (Section 7.4).

[6] If the linearization of the model is not sufficiently accurate, the quality of the evaluated uncertainty is affected, as is the estimate of the measurand. The latter point may be less well appreciated in some quarters. The bias so introduced into the estimate of the measurand is illustrated in Section 9.5, for example.

by a Gaussian or Student's-$t$ distribution with known mean and standard deviation and, in the latter case, a known number $\nu$ of degrees of freedom.

3. The adequacy of the Welch-Satterthwaite formula (cf. GUM Clause G.6.6, [37]).[7]

4. The input quantities to be uncorrelated if $\nu$ is finite.

### 5.4.2 When the conditions do or may not hold

In practice, Mainstream GUM is sometimes used in violation of the conditions listed in Section 5.4.1, and the results thus produced can only be regarded as approximate (with an unquantified degree of approximation). Or, more frequently, it is used without knowing whether these conditions hold (again with an unquantified degree of approximation). As indicated in Section 7, a *basic* form of MCS is readily implemented, requiring only model evaluation and simple random-number generation.[8] Because control can be exercised over the number of figures delivered (see Section 7.7.1), MCS can also be used to validate (i) the results provided by Mainstream GUM, and (ii) software implementations of Mainstream GUM. Although many evaluations based on Mainstream GUM may be sound, it is important to demonstrate that this is so. If (a legitimate implementation of) MCS indicated that certain Mainstream GUM results were invalid, it is recommended that consideration be given to using MCS instead.

### 5.4.3 Probability density functions or not?

The application of Mainstream GUM might not appear to require the specification of the input pdf's *per se*. It operates in terms of the means and standard deviations of these pdf's (Chapter 6). Mainstream GUM therefore has the apparent advantage that it is not necessary to provide the pdf's of the model inputs, i.e., just means and standard deviations would "suffice".

The Type A evaluations of the inputs are obtained by analyzing "repeated observations", from which means and standard deviations (but not pdf's) are obtained.

Conversely, for Type B evaluations, the means and standard deviations are determined from known or assigned input pdf's (Section 4.3). These pdf's are then used no further.

Thus, for some of the input quantities, the pdf's are not required and for the others they are not used. This attitude is seen as being incompatible with the Bayesian view that is increasingly used as a consistent basis for uncertainty evaluation. With a Bayesian approach, a pdf would be assigned to each input, based on whatever information, however meagre, is available.

As indicated in Section 6, the GUM in fact states (in Clause 4.1.6) that each input estimate and its associated standard uncertainty are obtained from a distribution of possible values of the input quantity $X_i$. Thus, a distribution is at least *implied*, although many practitioners would not obtain or even postulate

[7] The Welch-Satterthwaite formula is an approximation and assumes that the input mean values are independent and that the input standard deviations are independent.

[8] Implementations made by the authors have been applied to explicit and implicit models (where $Y$ can and cannot be expressed directly in terms of $X$), and complex-valued models (for electrical metrology), with univariate and multivariate outputs.

it, simply computing, for a Type A evaluation of uncertainty, an estimate and a standard deviation from the repeat observations to be used in the Mainstream GUM approach.

This guide encourages the assignment of a pdf to each input quantity. By so doing any of the candidate solution approaches considered in Section 5 can be applied. They can also be contrasted, if required. The assignment of these pdf's is addressed in Section 4.3.

## 5.5 Obtaining sensitivity coefficients

The sensitivity coefficients are the values of the partial derivatives of first order with respect to the input quantities, evaluated at the estimates of the input quantities. Their determination can present an algebraically difficult task. There are two stages:

1. Form algebraically the $n$ first-order partial derivatives,[9]

2. Evaluate these derivatives at the estimates of the input quantities.

These stages constitute Steps 3 and 5, respectively, of the Mainstream GUM procedure (as outlined in Section 5.2) for the calculation phase, Phase 2, of the uncertainty evaluation problem.

If the effort of determining these derivatives manually is considerable, there are two alternative approaches:

- Finite-difference methods,

- Computer-assisted algebraic methods.

Advice on the use of finite-difference methods is given in Section 5.5.1 and some comments on computer-assisted algebraic methods in Appendix C.

In the context of Phase 2, calculation, of the uncertainty evaluation problem there is no *essential* concept of sensitivity coefficients. They are of course required by the Mainstream GUM approach (Section 5.2). Independently of the approach used, they also convey valuable quantitative information about the influences of the various input quantities on the measurand (at least in cases where model linearization is justified). If an approach is used that does not require these coefficients for its operation, they can of course additionally be calculated if needed. Within the context of the Monte Carlo Simulation approach, it is also possible to apply the concept of sensitivity coefficients. Some comments are given in Appendix E.

### 5.5.1 Finite-difference methods

Numerical approximations to the values of derivatives can be obtained using finite-differences techniques. Given a value $i$ ($1 \leq i \leq n$), set all $X_\ell = x_\ell$, apart from $X_i$, i.e., assign the estimated values of the inputs to the input quantities, apart from the $i$th. Denote the resulting function of $X_i$ by $f_i(X_i)$.

[9] Expert advice may be required if the model is not continuously differentiable with respect to some or all of the input quantities.

A typical finite difference approximation to $\partial Y/\partial X_i$ evaluated at $\mathbf{x}$ is

$$\left.\frac{\partial Y}{\partial X_i}\right|_{\mathbf{X}=\mathbf{x}} \approx \frac{f_i(x_i + h_i) - f_i(x_i)}{h_i},$$

where $h_i$ is a "suitably small" increment in $x_i$ (see below). Note that $f_i(x_i) \equiv f(\mathbf{x})$ will already have been formed in evaluating the model at the estimates $\mathbf{x}$ of the input quantities.

The approximation can be perceived as follows. Consider the graph of $f_i(X_i)$. The formula gives the gradient of the chord joining the points $(x_i,\ f_i(x_i))$ and $(x_i + h_i,\ f_i(x_i + h_i))$. This gradient approximates the gradient of the tangent at $(x_i,\ f_i(x_i))$ to the graph of the function, which is of course the required derivative.

The choice of $h_i$ is important. If it is too great, the formula gives a large approximation error, i.e., the tangent and the chord point in appreciably different directions. If it is too small, the formula gives a large subtractive cancellation error, since the values of $f_i(x_i)$ and $f_i(x_i + h_i)$ will have many common leading figures.

A generally more accurate form, requiring an additional function evaluation, is

$$\left.\frac{\partial Y}{\partial X_i}\right|_{\mathbf{X}=\mathbf{x}} \approx \frac{f_i(x_i + h_i) - f_i(x_i - h_i)}{2h_i}.$$

For a given value of $h_i$, the magnitude of the approximation error is generally reduced using this form. Thus the value of $h_i$ can be larger, affording a better balance between approximation and cancellation errors.

The GUM, in Clause 5.1.3, suggests the use of the second formula with $h_i = u(x_i)$. This choice can generally be expected to be acceptable, although there may be exceptional circumstances.

# Chapter 6

# Mainstream GUM

## 6.1 Introduction to model classification

In the GUM [1] a measurement system is modelled by a functional relationship between the input quantities $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ and the measurand $Y$ (the output quantity) in the form

$$Y = f(\mathbf{X}). \tag{6.1}$$

In practice, however, this functional relationship does not apply *directly* for many of the measurement systems encountered, but may instead (a) take the form of an *implicit* relationship, $h(Y, \mathbf{X}) = 0$, (b) involve a *number of measurands* $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\mathrm{T}}$, or (c) involve *complex-valued quantities*. Although measurement models other than (6.1) are not directly considered in the GUM, the same underlying principles may be used to propagate uncertainties from the input quantities to the output quantities.

In Section 6.2 a classification of measurement models is given that is more general than that considered in the GUM. This classification is motivated by actual measurement systems, examples of which are given. For each measurement model it is indicated how the uncertainty of the measurement result is evaluated. Mathematical expressions for the uncertainty are stated using matrix-vector notation, rather than the subscripted summations given in the GUM, because generally such expressions are more compact and more naturally implemented within modern software packages and computer languages.

The Mainstream GUM approach is used throughout this chapter. Any doubt cocerning its applicability should be addressed as appropriate, for instance by using the concepts of Chapter 8.

## 6.2 Measurement models

A classification of measurement models is presented that depends on whether

1. There is one or more measurand, i.e., $Y$ is a scalar or a vector,

2. The measurand $Y$ is obtained by evaluating a formula or by solving an equation, or

3. The input quantities $\mathbf{X}$ are real- or complex-valued or the model function $f$ is real- or complex-valued or both $\mathbf{X}$ and $f$ are real- or complex-valued.[1]

The following information is assumed to be available:

1. Estimates $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ of the input quantities $\mathbf{X}$.
2. For $i = 1, \ldots, n$, *either*
   (a) The standard uncertainty $u(x_i)$, for uncorrelated input quantities, *or*
   (b) For $j = 1, \ldots, n$, the covariance $u(x_i,\ x_j)$ of $x_i$ and $x_j$, for correlated input quantities.[2] Note that $u(x_i,\ x_i) = u^2(x_i)$, the variance of $x_i$.

The following eight sub-sections provide matrix expressions for the uncertainty $u(y)$, in the form of the variance $u^2(y)$, of $y$ in the scalar case, or the covariance matrix $V_{\mathbf{y}}$ containing the covariances $u(y_i, y_j)$ in the vector case. Derivation of the formulae and equations is not given here. It is straightforward using basic statistical concepts and matrix manipulation.

The concentration is on providing information on the various types of model that appear in practice, and for each of these types giving relevant advice. The guidance is especially relevant when *software* is to be used to help provide uncertainty evaluations.[3]

For the first two model types (univariate, explicit, real-valued and multivariate, explicit, real-valued), the detail of the manner in which the matrices used are formed is provided through an example. The remaining model types are treated analogously.

### 6.2.1 Univariate, explicit, real-valued model

In a univariate, explicit, real-valued model, a single real-valued measurand $Y$ is related to a number of real-valued input quantities $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$ by an explicit functional relationship $f$ in the form of (4.2). This is the model directly considered in the GUM.

The measurement result is $y = f(\mathbf{x})$.

The (combined) standard uncertainty $u(y)$ of $y$ is evaluated from

$$u^2(y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} u(x_i, x_j), \tag{6.2}$$

where the partial derivatives[4] $\partial f/\partial x_i$ are referred to as *sensitivity coefficients*.

Write these covariances within the $n \times n$ matrix

$$V_{\mathbf{x}} = \begin{bmatrix} u(x_1, x_1) & \ldots & u(x_1, x_n) \\ \vdots & \vdots & \vdots \\ u(x_n, x_1) & \ldots & u(x_n, x_n) \end{bmatrix} \tag{6.3}$$

[1] If $\mathbf{X}$ or $f$ is complex-valued, the measurand $Y$ will in general also be complex-valued.

[2] Some or many of these covariance values may be zero.

[3] A supplemental guide to the GUM, based in part on the approach in this chapter, is being developed by JCGM/WG1.

[4] In accordance with the GUM, the informal notation that $\partial f/\partial x_i$ denotes the partial derivative $\partial f/\partial X_i$ evaluated at $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ is used. If derivatives are evaluated at any other point, a specific indication will be given.

and the sensitivity coefficients as the $1 \times n$ (row) vector

$$\nabla_{\mathbf{x}} f = (\partial f/\partial x_1 \ldots \partial f/\partial x_n). \tag{6.4}$$

Then, a compact way of writing (6.2), that avoids the use of doubly-scripted summations, is

$$u^2(y) = (\nabla_{\mathbf{x}} f) V_{\mathbf{x}} (\nabla_{\mathbf{x}} f)^{\mathrm{T}}, \tag{6.5}$$

where $V_{\mathbf{x}}$ is a matrix of order $n$ containing the covariance terms.

**Example 12** *End-gauge calibration*

[GUM Example H.1 End-gauge calibration] The length of a nominally 50 mm gauge block is determined by comparing it with a known gauge block standard of the same nominal length. An expression for the direct output of the comparison of the two gauge blocks is the difference

$$D = \{1 + (A_{\mathrm{S}} + \delta A)\Theta\} L - \{1 + A_{\mathrm{S}}(\Theta - \delta\Theta)\} L_{\mathrm{S}} \tag{6.6}$$

in their lengths, where[5]

- $L$ is the measurand, viz., the length at 20 °C of the gauge block being calibrated,
- $L_{\mathrm{S}}$ is the length of the standard at 20 °C as given in its calibration certificate,
- $A_{\mathrm{S}}$ is the coefficient of thermal expansion of the gauge block standard,
- $\delta A = A - A_{\mathrm{s}}$, where $A$ is the coefficient of thermal expansion of the gauge block being calibrated,
- $\Theta$ is the deviation in temperature from the 20 °C reference temperature of the gauge block being calibrated,
- $\delta\Theta = \Theta - \Theta_{\mathrm{S}}$, where $\Theta_{\mathrm{S}}$ is the deviation in temperature from the 20 °C reference temperature of the gauge block standard,

From (6.6) the measurand $L$ can immediately be expressed in terms of the quantities $D$, $L_{\mathrm{S}}$, $A_{\mathrm{S}}$, $\delta A$, $\Theta$ and $\delta\Theta$ as the model

$$L = \frac{\{1 + A_{\mathrm{S}}(\Theta - \delta\Theta)\} L_{\mathrm{S}} + D}{1 + (A_{\mathrm{S}} + \delta A)\Theta}.$$

In terms of the general formulation above, the input quantities are

$$\mathbf{X} \equiv (D, L_{\mathrm{S}}, A_{\mathrm{S}}, \delta A, \Theta, \delta\Theta)^{\mathrm{T}}$$

and the output quantity is

$$Y \equiv L.$$

The estimates of the input quantities are denoted by

$$\mathbf{x} \equiv (d, \ell_{\mathrm{S}}, \alpha_{\mathrm{S}}, \delta\alpha, \theta, \delta\theta)^{\mathrm{T}}. \tag{6.7}$$

[5]This choice of input variables is made for consistency with GUM, Example H.1. Other choices are possible. See later in this example.

The estimate

$$y \equiv \ell$$

of the measurand $L$ is

$$\ell = \frac{\{1 + \alpha_{\rm S}(\theta - \delta\theta)\}\,\ell_{\rm S} + d}{1 + (\alpha_{\rm S} + \delta\alpha)\theta}.$$

The partial derivatives of the model with respect to the input quantities are

$$\begin{aligned}
\frac{\partial L}{\partial D} &= \frac{1}{1 + (A_{\rm S} + \delta A)\Theta}, \\
\frac{\partial L}{\partial L_{\rm S}} &= \frac{1 + A_{\rm S}(\Theta - \delta\Theta)}{1 + (A_{\rm S} + \delta A)\Theta}, \\
\frac{\partial L}{\partial A_{\rm S}} &= \frac{(\Theta^2\delta A - \Theta\delta\Theta\delta A - \delta\Theta)L_{\rm S} - D\theta}{\{1 + (A_{\rm S} + \delta A)\Theta\}^2}, \\
\frac{\partial L}{\partial(\delta A)} &= -\frac{[\{1 + A_{\rm S}(\Theta - \delta\Theta)\}\,L_{\rm S} + D]\,\Theta}{\{1 + (A_{\rm S} + \delta A)\Theta\}^2}, \\
\frac{\partial L}{\partial\Theta} &= \frac{(A_{\rm S} + \delta A)(L_{\rm S}A_{\rm S}\delta\Theta - D) - L_{\rm S}\delta A}{\{1 + (A_{\rm S} + \delta A)\Theta\}^2}, \\
\frac{\partial L}{\partial(\delta\Theta)} &= \frac{-A_{\rm S}L_{\rm S}}{1 + (A_{\rm S} + \delta A)\Theta}.
\end{aligned}$$

The substitution ($d$ for $D$, $\ell_{\rm S}$ for $L_{\rm S}$, etc.) of the numerical values of the estimates (6.7) into these expressions for the partial derivatives yields the values of the sensivity coefficients.

The set of six sensitivity coefficients, arranged as a row vector, constitutes the row vector $\nabla_{\mathbf{x}} f$ in (6.5). The variances given by the squares of the standard uncertainties of the six input quantities constitute the diagonal elements of the covariance matrix $V_{\mathbf{x}}$ in (6.5). The remaining elements of $V_{\mathbf{x}}$ are taken as zero, since the input quantities are regarded as uncorrelated (GUM, Example H.1). Thus, $u^2(y)$ and hence $u(y)$ can be formed from (6.5).

∎

### 6.2.2 Multivariate, explicit, real-valued model

Although not directly considered in the *body* of the GUM, instances of measurement systems are included in that guide for which there is more than one measurand. This form of model occurs widely in metrological practice, viz., in calibration, instrument design and experimental data analysis.

Formally, the model for a multivariate, explicit, real-valued measurement system takes the form

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}), \tag{6.8}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\rm T}$ is a vector of $m$ measurands.

The measurement result is $\mathbf{y} = \mathbf{f}(\mathbf{x})$.

The uncertainty of $\mathbf{y}$ is expressed using a covariance matrix $V_{\mathbf{y}}$ that contains the covariances $u(y_i, y_j)$, and is evaluated by matrix multiplication from

$$V_{\mathbf{y}} = J_{\mathbf{x}} V_{\mathbf{x}} J_{\mathbf{x}}^{\rm T}, \tag{6.9}$$

where $J_{\mathbf{x}}$ is the $m \times n$ (Jacobian) matrix containing the values of the derivatives $\partial f_i / \partial x_j$, for $i = 1, \ldots, n$, $j = 1, \ldots, m$.

**Example 13** *Resistance and reactance of a circuit element*

The resistance $R$ and reactance $T$ of a circuit element are determined (GUM, Clause H.2) by measuring the amplitude $U$ of a sinusoidal alternating potential difference applied to it, the amplitude $I$ of the alternating current passed through it, and the phase shift angle $\phi$ between the two from

$$R = \frac{U}{I}\cos\phi, \qquad T = \frac{U}{I}\sin\phi.$$

In terms of the above notation, $\mathbf{X} \equiv (U,\ I,\ \phi)^{\mathrm{T}}$ and $\mathbf{Y} \equiv (R,\ T)^{\mathrm{T}}$.

The matrix $J_{\mathbf{x}}$, of dimension $2 \times 3$, is

$$J_{\mathbf{x}} = \begin{bmatrix} \frac{\partial R}{\partial U} & \frac{\partial R}{\partial I} & \frac{\partial R}{\partial \phi} \\ \frac{\partial T}{\partial U} & \frac{\partial T}{\partial I} & \frac{\partial T}{\partial \phi} \end{bmatrix} = \begin{bmatrix} \frac{1}{I}\cos\phi & -\frac{U}{I^2}cos\phi & -\frac{U}{I}\sin\phi \\ \frac{1}{I}\sin\phi & -\frac{U}{I^2}\cos\phi & \frac{U}{I}\cos\phi \end{bmatrix}.$$

The substitution of estimates ($u$ of $U$, etc.) into the expression for $J_{\mathbf{x}}$ will yield this matrix, and, given the covariance matrix $V_{\mathbf{x}}$ of order 3 of the input quantities (cf. Section 6.2.1), the covariance matrix $V_{\mathbf{y}}$, of order 2 of the output quantities is given by matrix multiplication using (6.9).

∎

### 6.2.3 Univariate, implicit, real-valued model

In a univariate, implicit, real-valued model, a single real-valued measurand $Y$ is related to real-valued input quantities $\mathbf{X}$ in a way that cannot readily or stably be represented by an explicit function. A model for the measurement system takes the form of an *equation* relating $\mathbf{X}$ and $Y$:

$$h(Y,\ \mathbf{X}) = 0. \tag{6.10}$$

The measurement result $y$ is given by the solution of the equation $h(y,\ \mathbf{x}) = 0$. This equation is solved numerically for $y$ using a zero-finding algorithm [25, 34], such as the bisection algorithm in a case where suitable lower and upper bounds for $y$ are known. The standard uncertainty $u(y)$ of $y$ is evaluated from

$$u^2(y)\left(\frac{\partial h}{\partial y}\right)^2 = (\nabla_{\mathbf{x}} h) V_{\mathbf{x}} (\nabla_{\mathbf{x}} h)^{\mathrm{T}}, \tag{6.11}$$

where $\nabla_{\mathbf{x}} h$ is the row vector of sensitivity coefficients of $h$ with respect to $\mathbf{X}$, evaluated at $\mathbf{x}$ (cf. (6.4)).

**Example 14** *The pressure generated by a pressure balance*

The pressure $p$ generated by a pressure balance is defined implicitly by the equation[6]

$$p = \frac{M(1 - \rho_a/\rho_m)g_\ell}{A_0(1 + \lambda p)(1 + \alpha(T - 20))}, \tag{6.12}$$

[6] More complete models can also be considered [47] that include, for example, a correction to account for surface tension effects.

where $M$ is the total applied mass, $\rho_a$ and $\rho_m$ are the densities of air and the applied masses, $g_\ell$ is the local acceleration due to gravity, $A_0$ is the effective cross-sectional area of the balance at zero pressure, $\lambda$ is the distortion coefficient of the piston-cylinder assembly, $\alpha$ is the temperature coefficient, and $T$ is temperature [47].

There are eight input quantities, $\mathbf{X} \equiv (A_0, \lambda, \alpha, T, M, \rho_a, \rho_m, g_\ell)^{\mathrm{T}}$ and a single measurand $Y \equiv p$ related by the implicit model[7]

$$h(y,\ \mathbf{x}) = A_0 p(1 + \lambda p)(1 + \alpha(T - 20)) - M(1 - \rho_a/\rho_m)g_\ell = 0. \tag{6.13}$$

Given estimates $\mathbf{x}$ of $\mathbf{X}$, Equation (6.13) is solved for $y \equiv p$. The first-order partial derivatives of $h$, in (6.13), with respect to $\mathbf{X}$, evaluated at $\mathbf{x}$, provide the elements of the row vector $\nabla_{\mathbf{x}} h$ in (6.11). Together with the covariance matrix $V_{\mathbf{x}}$ of the estimates $\mathbf{x}$ and the partial derivative $\partial h/\partial y \equiv \partial h/\partial p$, evaluated at $\mathbf{x}$, this information permits $u(y) \equiv u(p)$ to be formed using (6.11).

### 6.2.4 Multivariate, implicit, real-valued model

A multivariate, implicit, real-valued model is identical to (6.10), but $Y$ is now a vector, in the form of a vector measurand $\mathbf{Y}$:

$$\mathbf{h}(\mathbf{Y},\ \mathbf{X}) = \mathbf{0}. \tag{6.14}$$

The measurement result $\mathbf{y}$ is given by the solution of the equation $\mathbf{h}(\mathbf{y};\ \mathbf{x}) = \mathbf{0}$. These equations are solved numerically for $\mathbf{y}$ using an iterative algorithm such as Newton's method [34], starting from a specified point $\mathbf{y}^{(0)}$. The covariance matrix $V_{\mathbf{y}}$ for $\mathbf{y}$ is related to that, $V_{\mathbf{x}}$, for $\mathbf{x}$ by

$$J_{\mathbf{y}} V_{\mathbf{y}} J_{\mathbf{y}}^{\mathrm{T}} = J_{\mathbf{x}} V_{\mathbf{x}} J_{\mathbf{x}}^{\mathrm{T}}, \tag{6.15}$$

a system of linear equations that is solved for $V_{\mathbf{y}}$.[8]

[7] There is not a unique way to write the implicit model. For instance, in place of (6.13) the model given by the difference between the left- and right-hand sides of (6.12) could be used. The efficiency and stability of the solution of the model equation depends on the choice made.

[8] Using recognised concepts from numerical linear algebra [35], a numerically stable way to form $V_{\mathbf{y}}$, that accounts for the fact that $J_{\mathbf{x}}$ is a rectangular matrix and $J_{\mathbf{y}}$ a square matrix, is as follows:

1. Form the Cholesky factor $R_{\mathbf{x}}$ of $V_{\mathbf{x}}$, i.e., the upper triangular matrix such that $R_{\mathbf{x}}^{\mathrm{T}} R_{\mathbf{x}} = V_{\mathbf{x}}$.
2. Factorize $J_{\mathbf{x}}$ as the product $J_{\mathbf{x}} = Q_{\mathbf{x}} U_{\mathbf{x}}$, where $Q_{\mathbf{x}}$ is an orthogonal matrix and $U_{\mathbf{x}}$ is upper triangular.
3. Factorize $J_{\mathbf{y}}$ as the product $J_{\mathbf{y}} = L_{\mathbf{y}} U_{\mathbf{y}}$, where $L_{\mathbf{y}}$ is lower triangular and $U_{\mathbf{y}}$ is upper triangular.
4. Solve the matrix equation $U_{\mathbf{y}}^{\mathrm{T}} M_1 = I$ for $M_1$.
5. Solve $L_{\mathbf{y}}^{\mathrm{T}} M_2 = M_1$ for $M_2$,
6. Form $M_3 = Q_{\mathbf{x}}^{\mathrm{T}} M_2$.
7. Form $M_4 = U_{\mathbf{x}}^{\mathrm{T}} M_3$.
8. Form $M = R_{\mathbf{x}} M_4$.
9. Orthogonally triangularize $M$ to give the upper triangular matrix $R$.
10. Form $V_{\mathbf{y}} = R^{\mathrm{T}} R$.

It is straightforward to verify this procedure using elementary matrix algebra.

**Example 15** *Correlated pressures generated by a pressure balance*

In the example of Section 6.2.3, let $p_i$, $i = 1, \ldots, m$, denote the generated pressures for applied masses $M_i$ and temperatures $T_i$, with $A_0$, $\lambda$, $\alpha$, $\rho_a$, $\rho_m$ and $g_\ell$ fixed. Each $p_i$ is obtained by solving an equation of the form (6.13). However, the resulting values are not independent because they all depend on the measured quantities $A_0$, $\lambda$, $\alpha$, $\rho_a$, $\rho_m$ and $g_\ell$. To understand the correlation between the $p_i$, it is necessary to model the system using a multivariate implicit function in which $\mathbf{X} \equiv (A_0, \lambda, \alpha, T_1, M_1, \ldots, T_m, M_m, \rho_a, \rho_m, g_\ell)^{\mathrm{T}}$ is the vector of input quantities and $\mathbf{Y} \equiv (p_1, \ldots, p_m)^{\mathrm{T}}$ the vector of measurands.

### 6.2.5 Univariate, explicit, complex-valued model

Let $x$ be a complex-valued quantity written in terms of its real and imaginary parts:

$$x = x_R + \sqrt{-1}x_I.$$

The uncertainty of $x$ is expressed using a $2 \times 2$ matrix

$$V = \begin{bmatrix} u^2(x_R) & u(x_R, x_I) \\ u(x_R, x_I) & u^2(x_I) \end{bmatrix}.$$

This is a more complicated data structure than for the case of real-valued $x$.[9] For an $n$-vector $\mathbf{x}$ of complex-valued quantities, the uncertainty of $\mathbf{x}$ is expressed using a $2n \times 2n$ matrix $V_{\mathbf{x}}$:

$$V_{\mathbf{x}} = \begin{bmatrix} V_{1,1} & \cdots & V_{1,n} \\ \vdots & \ddots & \vdots \\ V_{n,1} & \cdots & V_{n,n} \end{bmatrix}, \tag{6.16}$$

where $V_{i,i}$ is a $2 \times 2$ sub-matrix containing the uncertainty of $x_i$, and $V_{i,j}$, $i \neq j$, is a $2 \times 2$ sub-matrix of $V_{\mathbf{x}}$ containing the covariances of the real and imaginary parts of $x_i$ with those of $x_j$.

In a univariate, explicit, complex-valued model, a single complex-valued measurand $Y$ is related to a number of complex-valued input quantities $\mathbf{X}$ by an explicit functional relationship in the form of (6.1). The uncertainty $V_y$ of the measurement result $y$ is evaluated from

$$V_y = J_{\mathbf{x}} V_{\mathbf{x}} J_{\mathbf{x}}^{\mathrm{T}}, \tag{6.17}$$

where $J_{\mathbf{x}}$ is a $2 \times 2n$ matrix whose first row contains the derivatives of the real part of $f$ with respect to the real and imaginary parts of $\mathbf{X}$, and in whose second row are the derivatives of the imaginary part of $f$.

**Example 16** *The reflection coefficient measured by a calibrated microwave reflectometer*

[9] The data structure is, however, like that for the vector $\mathbf{X} = (X_R,\ X_I)^{\mathrm{T}}$.

The (complex-valued) reflection coefficient $\Gamma$ measured by a calibrated microwave reflectometer, such as an automatic network analyser (ANA), is given by

$$\Gamma = \frac{aW + b}{cW + 1}, \tag{6.18}$$

where $W$ is the observed (complex-valued) uncorrected reflection coefficient and $a$, $b$ and $c$ are (complex-valued) calibration coefficients characterizing the reflectometer [31, 45, 60].

### 6.2.6 Multivariate, explicit, complex-valued model

In a multivariate, explicit, complex-valued model the measurement system model (6.8) applies with $\mathbf{X}$ and $\mathbf{Y}$ complex-valued. The uncertainty of $\mathbf{y}$ is given by the $2n \times 2m$ matrix $V_{\mathbf{y}}$ (see (6.16)) evaluated from

$$V_{\mathbf{y}} = J_{\mathbf{x}} V_{\mathbf{x}} J_{\mathbf{x}}^{\mathrm{T}},$$

where $J_{\mathbf{x}}$ is a $2m \times 2m$ matrix containing the derivatives of the real and imaginary parts of each component of $\mathbf{f}$ with respect to the real and imaginary parts of each component of $\mathbf{X}$.

**Example 17** *A calibrated microwave reflectometer used to measure correlated reflection coefficients*

Let $a$, $b$ and $c$ be the calibration coefficients for an ANA as in the example of Section 6.2.5. Suppose $W_i$, $i = 1, \ldots, m$, are $m$ observed uncorrected reflection coefficients for which the corresponding "true" reflection coefficients are $\Gamma_i$, $i = 1, \ldots, m$. The values $\Gamma_i$ are obtained by evaluating $m$ formulae of the form of (6.18). However, the measurement results $\Gamma_i$ are not independent because they all depend on the calibration coefficients. To understand the correlation between the $\Gamma_i$, it is necessary to model the system using a multivariate explicit function in which the vector of input quantities $\mathbf{X} \equiv (a, b, c, W_1, \ldots, W_m)^{\mathrm{T}}$ and the vector measurand $\mathbf{Y} \equiv (\Gamma_1, \ldots, \Gamma_m)^{\mathrm{T}}$.

### 6.2.7 Univariate, implicit, complex-valued model

In a univariate, implicit, complex-valued model, the measurement model (6.10) applies with $Y$ and $\mathbf{X}$ complex-valued. The uncertainty of $y$ is evaluated from

$$J_y V_y J_y^{\mathrm{T}} = J_{\mathbf{x}} V_{\mathbf{x}} J_{\mathbf{x}}^{\mathrm{T}},$$

where $J_y$ is a $2 \times 2$ matrix containing the derivatives of the real and imaginary parts of $h$ with respect to the real and imaginary parts of $y$. Compare with Section 6.2.4

**Example 18** *The reflection coefficient measured by a calibrated microwave reflectometer*

Another approach to the example given in Section 6.2.5 is to relate the inputs $\mathbf{X} \equiv (a, b, c, W)^{\mathrm{T}}$ and the measurand $Y \equiv \Gamma$ using the (complex-valued) implicit model

$$h(Y, \mathbf{X}) = \Gamma(cW + 1) - (aW + b) = 0.$$

An advantage of this approach is that the calculation of derivatives and thence sensitivity coefficients is easier.

### 6.2.8 Multivariate, implicit, complex-valued model

In a multivariate, implicit, complex-valued model, the measurement system model (6.14) applies with $\mathbf{X}$ and $\mathbf{Y}$ complex-valued. The uncertainty of $\mathbf{y}$ is then evaluated from (6.15) which constitutes a linear system for $V_{\mathbf{y}}$.

**Example 19** *Calibration of a microwave reflectometer using three standards*

The calibration of a reflectometer is undertaken by measuring values $W$ corresponding to a number of standards $\Gamma$. Typically, three standards are used, giving the three simultaneous equations

$$\Gamma_i(cW_i + 1) - (aW_i + b) = 0, \qquad i = 1, 2, 3,$$

that are solved for the three calibration coefficients $a$, $b$ and $c$. There are six (complex-valued) input quantities $\mathbf{X} \equiv (W_1, \Gamma_1, W_2, \Gamma_2, W_3, \Gamma_3)^{\mathrm{T}}$ and three (complex-valued) measurands $\mathbf{Y} \equiv (a, b, c)^{\mathrm{T}}$ related by a model of the form (6.14), where

$$h_i(\mathbf{y}, \mathbf{x}) = \Gamma_i(cW_i + 1) - (aW_i + b) = 0, \qquad i = 1, 2, 3.$$

## 6.3 Classification summary

The above classification of measurement system models is more general than that considered in Mainstream GUM. The classification is motivated by actual measurement systems for which the guidance provided in the GUM is applicable but not immediately so.

# Chapter 7

# Monte Carlo Simulation

The manner in which a general *numerical* approach, Monte Carlo Simulation (MCS), can be applied to uncertainty evaluation is described. Practical implementation considerations are given.[1]

In the context of uncertainty evaluation, MCS is a *sampling technique* that provides an alternative approach to the propagation of uncertainties: the process is undertaken numerically rather than analytically. The technique is also useful for validating the results returned by the application of Mainstream GUM (Section 6), as well as in circumstances where the assumptions made by Mainstream GUM may not apply. In fact, it provides much richer information, by propagating the pdf's (rather than just the uncertainties) of the input quantities $\mathbf{X}$ through the measurement model $f$ to provide the pdf of the measurand $Y$ or the joint pdf's of multivariate measurands $\mathbf{Y}$. From the pdf of the output quantity coverage intervals (in the univariate case) can then straightforwardly be produced, as can other statistical information.[2]

MCS enables account to be taken of the knowledge of analytically or experimentally specified input pdf's, including asymmetric densities such as Poisson (counting rates) and Gamma (special cases of which are exponential and chi-squared). The pdf's for the input quantities form the necessary basis for determining the pdf of an output quantity by MCS. (A calculated mean and standard deviation, as provided by Mainstream GUM, do not alone form such a basis.)

Experimentally specified pdf's arise especially as the output from a previous application of MCS in a process having several stages, where the outputs from one stage becomes input to the next (Section 4.2.3). Re-sampling from these results is equivalent to the use of bootstrapping, in which no prior pdf of a quantity is assumed and the posterior distribution is constructed by assigning the probability $1/M$ to each of $M$ observed values of the quantity and zero probability to all other values. This approach is consistent with the use of an empirical output pdf, generated by MCS, to estimate the distribution function and hence coverage intervals.

If the model inputs have correlated errors, sampling would take place from

[1] A supplemental guide to the GUM, related to the approach in this chapter, is being developed by JCGM/WG1.

[2] The determination of coverage *regions* (for the multivariate case) remains a research problem. See Section 7.4.

the corresponding joint pdf's. A general approach to such sampling is available [64].

MCS is a step-by-step procedure, like Mainstream GUM. The difference is that in MCS a small number of steps is repeated very many times, and the results obtained aggregated. Hence, computer implementation is essential. Increasing use of software is being made in applying Mainstream GUM, so the use of software for MCS is seen as a practical and acceptable (and more general) alternative. Specifications for key software units are available [20].

The technique used is that of repeated *sampling* from the pdf's describing the input quantities. The fact that the sampling is carried out from the provided pdf's rather than being based on approximations the quality of which is difficult to quantify is seen as highly relevant in removing the influence of uncontrollable limitations.

So, given the model and the pdf's of its input quantities, Monte Carlo Simulation constitutes a tool to estimate the pdf of the scalar output quantity $Y$ or vector output quantity $\mathbf{Y}$.

The output pdf is fundamental to determining any or all statistics associated with the measurement result.[3] From it can be obtained:

- Mean[4], median, mode and other estimates of location such as the total median [24],
- Standard deviation (standard uncertainty) and its square, the variance, and higher moments such as skewness and kurtosis,[5]
- A coverage interval at some stated level of probability (the generalization of "measurement result $\pm$ expanded uncertainty" in the case of a Gaussian output),

[3]Recall that the output quantity may become the input quantity to a subsequent stage in a multi-stage model (Section 4.2.3), and hence in these circumstances MCS provides valuable problem-specific information that would not necessarily be provided by more traditional approaches to uncertainty evaluation.

[4]There is currently a debate in the metrology community concerning whether *this* value or the value of the model corresponding to the estimates of the input quantities should be used. In many instances it will make neglible practical difference. In some cases, the difference can be appreciable. Consider the simple model $Y = X_1^2$, where the pdf assigned to $X_1$ has mean zero and standard uncertainty $u$. Taking the mean of values of $Y$ involves averaging a set of non-negative values and hence will be positive. In contrast, the value of $Y$ corresponding to the mean value of $X$ is zero. In this case, the former value is more reasonable, since zero lies at an extremity of the range of possible values for the output quantity and is hardly representative, as a mean should be. In other, somewhat more complicated situations, the mean of the values of the output quantity constitutes a measurement result that contains unwanted bias. In this circumstance, it can be more meaningful to take instead the value of $Y$ corresponding to the mean value of $X$. In general, circumstances should dictate the choice. In one sense the degree of arbitrariness is genuine. The Monte Carlo procedure naturally provides *fractiles* of the distribution function of the output quantity. In particular the 0.025 and 0.975 fractiles define a 95% coverage interval for the measurand. Such a coverage interval is also given by any other pair of fractiles that differ by 0.95, such as 0.015 and 0.965, or 0.040 and 0.990. In this setting, it is less meaningful to quote a coverage interval in the form $y \pm U(y)$, involving the "mean" $y$. For comparison with a "Monte Carlo" interval, it would instead be appropriate to quote the interval $[y - U(y),\ y + U(y)]$.

[5]The first moment of a pdf is the mean, a measure of location, the second is the variance, a measure of dispersion or spread about the mean, the third is skewness, a measure of asymmetry about the mean, and the fourth is kurtosis, a measure of heaviness of the tails of the pdf or the peakedness in the centre. [51, p143, p329]

- Any other statistical estimator or derived statistical quantity.

For multivariate output quantities, there would be higher-dimensional counterparts of these quantities.

The use of MCS is in principle straightforward, but a solid implementation requires (a) generators (algorithms) to sample from all (joint) input pdf's likely to be useful in practice (some of which will be multidimensional because of correlations) and (b) consideration of the number of Monte Carlo trials needed to deliver two (say) correct figures in the output uncertainty. Work is needed on (a) to cater for the variety of possible input pdf's. As indicated, a general approach to such sampling is available [64]. This and other approaches need to be reviewed carefully for their suitability in the current context. For some of the commonest pdf's (univariate uniform, univariate Gaussian and multivariate Gaussian), generators to carry out the sampling are available (Section 7.7). For (b), see Section 7.7.1. The software specifications [20] are relevant here.

The MCS sampling technique is also valuable for validating the results returned by the application of Mainstream GUM, as well as in circumstances where the assumptions made by Mainstream GUM do not apply (Section 8). Further, the fact that MCS permits general pdf's rather than just means and uncertainties to be propagated through measurement-system models cannot be underestimated. As indicated, all statistical information relating to the variability of the results of measurements, including correlation effects, can be discerned from these output distributions. The quality of this information will depend on that of the model and the input quantities and, if those are considered acceptable, is only limited by the number of samples generated. In particular, quantitative results relating to the "propagation of uncertainties" can be obtained from the propagated pdf's. In contrast, the converse is not true: the propagation of uncertainties as, e.g., in Mainstream GUM, cannot be used to derive the output pdf's (unless it can be shown that it is acceptable to regard them as Gaussian).

The Mainstream GUM approach to uncertainty evaluation is based on propagating uncertainties in a first-order approximation to the model of the measurement system. MCS sampling techniques [29, 30] provide an alternative approach in which instead the statistical distributions are propagated. Although no such approximation as above is made, e.g., the nonlinearity of the model is taken into account, the sampling process introduces a sampling error that depends on the number of samples chosen or available.

A major distinction is that with Mainstream GUM there is no control over the extent of the approximation introduced by linearizing the model, or assuming the output distribution function is Gaussian, whereas with MCS the sampling error can be influenced by the amount of sampling, the number of Monte Carlo trials (Section 7.7.1).

## 7.1 Sampling the input quantities

This section is concerned with the manner in which the inputs to the model can be sampled as part of the Monte Carlo Simulation process. There are two main types of input: a pdf that is defined continuously and a pdf that is defined discretely. Moreover, a pdf can be univariate (Section 7.1.1) or multivariate (joint) (Section 7.1.2).

### 7.1.1 Univariate probability density functions

Each independent input quantity has an associated pdf (GUM Clause G.1.4). Consider first a pdf defined continuously, e.g., a uniform, Gaussian or Student's-$t$ pdf. Sampling from such a pdf is carried out using a (pseudo)random number generator that is appropriate for that pdf. See Section 7.7.

Consider now an input quantity defined by a sufficiently large number of repeated observations, and for which nothing further is known about the input. In such a case the observations themselves can be used to represent their distribution.

Consider such an input variable $q$ specified by $n_q$ observations $q_1, \ldots, q_{n_q}$. Let $q_{(1)}, \ldots, q_{(n_q)}$ denote these values arranged in non-decreasing order. The join $Q(q)$ of the points $(q_{(i)},\ (i-1/2)/n_q)$, $i = 1, \ldots, n_q$, provides an estimate of the distribution function (the indefinite integral of the pdf) of $q$.

Sampling from this function can be carried out using a uniform generator and inverse linear interpolation. Specifically, a sample is given by

1. Using a generator to provide a value from the uniform pdf defined over zero to one,
2. Using inverse linear interpolation to provide a value of $q$ satisfying $Q(q) = z$.

This procedure is not entirely satisfactory since for $(1/n_q)$th of the time, on average, a value returned by the uniform generator will lie beyond the join $Q(q)$, i.e., it will lie in a tail of the pdf, viz., $z < 1/(2n_q)$ or $z > 1-1/(2n_q)$. This failure to sample from the tails of the pdf can be addressed by introducing an heuristic. If $z < 1/(2n_q)$ or $z > 1 - 1/(2n_q)$, a new sample is drawn from the uniform distribution (and re-drawn as many times as necessary). Alternatively, a value of $z < 1/(2n_q)$ can be replaced by this limit, and similarly when $z > 1 - 1/(2n_q)$, a form of Winsorising. Both of these (and other such) approaches will bias the solution, since no tail information (at least beyond the above limits) is sampled. The latter is more appropriate than the former, however, because an "unacceptable" value is replaced by the nearest "acceptable" value rather than by a freshly sampled value. The use of parametric information, if available, would help, but would be against the statement that only the data itself is available.

Particularly for a large sample, however, the effect of tail information on a consequent coverage interval is not appreciable. Large samples are therefore strongly recommended.

Alternative, and more commonly, the *bootstrap* can be used. The bootstrap simply draws a uniformly random observation from the set. A value is "replaced" after being drawn, so that all observations maintain the same probability of being drawn.

### 7.1.2 Multivariate probability density functions

Sampling from a joint pdf that is defined continuously is largely a research topic. Advice should be sought from a statistician or numerical analyst. A joint (multivariate) Gaussian distribution can straightforwardly be handled [20].

Sampling from a joint pdf that is defined by a set of discrete values can be carried out straightforwardly. Such values are likely to have been obtained from

a previous application of MCS to a multivariate model (Section 7.4). Suppose $M$ Monte Carlo trials have previously be carried out and $n$ is the number of output quantities in that simulation. This information will have been preserved as an $n \times M$ array of values. It embodies the full correlation effects present in those outputs. A column chosen at random from the array will be a valid sample with its correlation effects.

## 7.2 Univariate models

Consider first the *univariate*[6] model function

$$Y = f(\mathbf{X}),$$

where

$$\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}.$$

Let the probability density function (pdf) of the $i$th input quantity $X_i$ be $g_i(X_i)$ and the pdf of $Y$ be $g(Y)$. Let

$$G(Y) = \int_{-\infty}^{Y} g(t)dt$$

denote the *distribution function* corresponding to $g$.

Adequate estimation of $G(Y)$ will permit all the required statistics associated with $Y$ to be determined.

## 7.3 A simple implementation of MCS for univariate models

Advice on a simple implementation of MCS is given in the case of the univariate model, above. Its use will permit for instance the estimation of the (combined) standard uncertainty of the measurement result $y$, and a 95% coverage interval for the measurand $Y$.

The number $M$ of Monte Carlo trials is to be determined. In the experience of the authors a value of the order of 50,000 is often appropriate. However, guidance on taking a number of trials to provide a prescribed accuracy in the estimated uncertainties is given in Section 7.7.1. It is recommended that such guidance be applied rather than selecting a *fixed* number (e.g., 50,000) *a priori*, even though such a number is often reasonable.[7] Those organisations that have a number of similar uncertainty calculations to make can reasonably infer a value for $M$ by noting the value for a representative problem, and using that. The computational penalty for employing the adaptive procedure of Section 7.7.1 is so small, however, that it is recommended to use it.

The procedure is as follows.

[6] A univariate model function (Section 6.2.1, e.g.) has a single (scalar) output $Y$ and an arbitrary number $n$ of inputs $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$. A multivariate model function (Section 6.2.3, e.g.) has an arbitrary numbers of inputs *and* outputs. The latter is considered in Section 7.4.

[7] The number of trials needed is a sensitive function of the level of probability required. The "typical" figure of 50,000 applies to a 95% level of probability for a "moderate" problem. Because of the need to quantify the tails of the output pdf sufficiently well, very many more trials will be needed for a higher level such as 99%.

1. Select a number $M$ of Monte Carlo trials to be made.
2. Generate $M$ samples $\mathbf{x}_i$ of the input quantities $\mathbf{X}$.
3. For each $\mathbf{x}_i$, evaluate the model to give the values
$$y_i = f(\mathbf{x}_i), \qquad i = 1, \ldots, M,$$
of the output quantity.
4. Regard these values of $y_i$ when assembled into a histogram (with suitable cell widths) as a (scaled) empirical estimate of the pdf of $Y$.[8]
5. Take the mean of these values of $y_i$ as the measurement result $y$.
6. Take the standard deviation of the $y_i$ as the (combined) standard uncertainty $u(y)$ of $y$.
7. Sort the values $y_i$, $i = 1, \ldots, M$, into non-decreasing order. Denote the sorted values by $y_{(i)}$, $i = 1, \ldots, M$.
8. Form the join of the points $(y_{(i)}, p_i)$, $i = 1, \ldots, M$, where $p_i = (i-1/2)/M$. It provides an empirical estimate of the distribution function of $Y$.
9. Form the 2.5-and 97.5-percentiles $y_{(0.025M)}$ and $y_{(0.975M)}$.
10. Take $(y_{(0.025M)}, y_{(0.975M)})$ as a 95% coverage interval for the output.[9]

The procedure depends on the ability to sample from the pdf's of the input quantities. Section 7.7 provides advice on this aspect.

In the procedure, the number $M$ of Monte Carlo trials is selected initially, at Step 1. The variant, indicated above in this section, in which the number is chosen adaptively, i.e., as the procedure is followed, is given in Section 7.7.1.

**Example 20** *The formation of the endpoints of a coverage interval from a discrete estimate of a distribution function*

For $M = 50,000$, since $0.025M = 1,250$ and $0.975M = 48,750$, a 95% coverage interval is taken as $(y_{(1250)}, y_{(48750)})$, the values in positions 1250 and 48750 in the ordered list of values $y_{(1)}, \ldots, y_{(50000)}$.



**Example 21** *The use of Monte Carlo Simulation: estimation of the distribution function for a simple nonlinear model*

[8] Most subsequent calculations will not be carried out in terms of this histogram, the "shape" of which depends sensitively on the choice of cell size [33], but in terms of the distribution function. The histogram is, however, a useful visual aid to understanding the nature of the pdf.

[9] The $100p$-percentile is the $(Mp)$th value in the list. If $Mp$ is not an integer, the integer immediately below $Mp$ is taken, if $p < 1/2$, and the integer immediately above $Mp$, otherwise. Similar considerations apply in other circumstances.

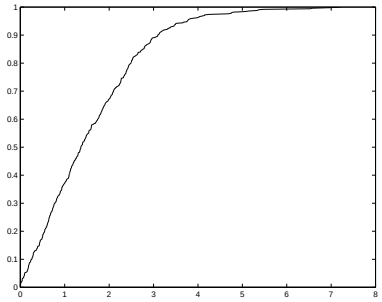

Figure 7.1: An empirical estimate obtained using Monte Carlo Simulation, with 500 trials, of the distribution function for the model $Y = X_1^2$. $X_1$ has a probability density function that is Gaussian with mean 1.2 and standard deviation 0.5.

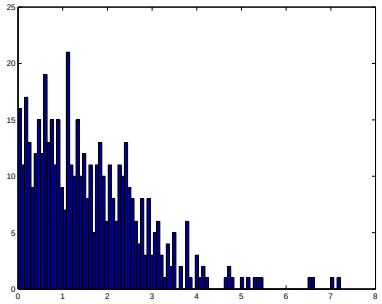

Figure 7.2: A histogram of the values used to produce the distribution function of Figure 7.1. It constitutes a discrete, scaled representation of the corresponding probability density function.

Consider the univariate model $Y = X_1^2$, where the single input $X_1$ has a Gaussian pdf with mean 1.2 and standard deviation 0.5. Determine the pdf and distribution function of $Y$, the mean and standard deviation of $Y$ and a 95% coverage interval for the measurand.

First, the number $M$ of Monte Carlo trials was taken as 500. Values $x_i$, $i = 1, \ldots, M$, were sampled from the Gaussian distribution having mean 1.2 and standard deviation 0.5. The corresponding values $y_i = x_i^2$, $i = 1, \ldots, M$, were calculated according to the model. The empirical distribution function of $Y$ was taken, in accordance with the above, as the join of the points $(y_{(i)}, p_i)$, $i = 1, \ldots, M$, where $p_i = (i - 1/2)/M$. Figure 7.1 shows the distribution function so obtained.

A histogram of the values of $y_i$ appears as Figure 7.2. It constitutes a discrete, scaled representation of the pdf.

The empirical distribution function is a much smoother function than the empirical pdf, the main reason for which is that the empirical distribution function constitutes a cumulative sum of the empirical pdf. Taking successive sums is a smoothing process, being the converse of taking successive differences. This summation corresponds to the integration counterpart for analytical pdf's.

The exercise was repeated for $M = 50,000$ trials. See Figures 7.3 and 7.4.

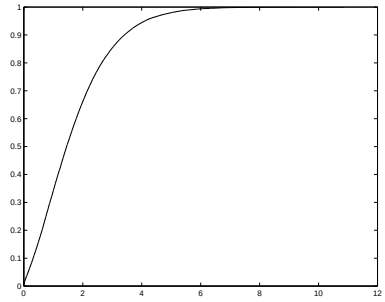

Figure 7.3: As Figure 7.1 but based on 50,000 Monte Carlo trials.

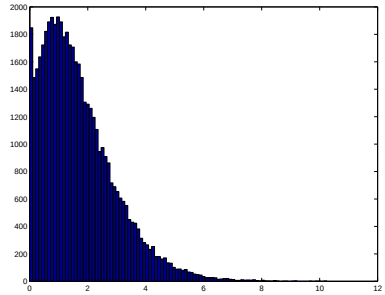

Figure 7.4: As Figure 7.2 but based on 50,000 Monte Carlo trials.

The enhanced smoothness of the results is evident. Statistics computed from the larger sample would be much more reliable. It can be expected that increasing the number of trials by a factor of 100, as here, would yield an additional significant digit of accuracy in the computed statistics [19]. A rule for the number of trials related to this consideration is given in Section 7.7.1.

The enhanced resolution permits a feature to be discerned in the pdf for $M = 50,000$ (Figure 7.4) that was not evident in that for $M = 500$. The pdf is *bimodal*, there being a narrow peak near the origin, in addition to the main peak. This is not an artifact introduced by the sampling procedure, but a genuine effect. Its presence is due to the fact that 0.8% of the values of $X_1$ according to its pdf are negative. These values lie in the left-hand tail of the Gaussian pdf for $X_1$, i.e., that with mean $\mu = 1.2$ and standard deviation $\sigma = 0.5$. The above proportion corresponds to the area under the standardized Gaussian curve (i.e., that with mean zero and standard deviation unity) to the left of the value $z = (0 - \mu)/\sigma = -2.4$. These values when squared, through the model $Y = X_1^2$, are aggregated with those arising from small positive values of $X_1$. Even for such a superficially innocent example, there can be a "surprise" such as this!

The mean and standard deviation of the output as estimated by Mainstream GUM are 1.44 and 1.20. Those provided by MCS were 1.70 and 1.26. The standard deviation in this example is reasonably estimated by Mainstream GUM, but the mean is lower than the correct value.

A further noteworthy feature arises from this example. In the case $M = 50,000$, a 95% coverage interval for $Y$, determined from the 2.5- and 97.5-

percentiles of the empirical distribution function was [0.1, 4.8]. That provided using Mainstream GUM is $[-1.1,\ 3.9]$ or, equivalently, $1.4 \pm 2.5$. The lengths of the coverage interval are similar. However, the interval provided by Mainstream GUM is shifted left relative to that for MCS. In fact, the portion of the Mainstream GUM coverage interval from -0.8 to zero is infeasible, since, from its definition, $Y$ cannot be negative.

Coverage intervals at other levels of probability were also obtained using MCS and Mainstream GUM. Appreciable differences were again observed. For instance, at a 99.8% level of probability (corresponding to a coverage factor of 3 under the Gaussian assumption), the coverage interval provided by MCS was $[0.0,\ 7.5]$ and that for Mainstream GUM is $[-2.3,\ 5.2]$.

∎

Although the above example might seem extreme, situations with large uncertainties arise in metrology areas such as EMC measurement. Instances where the quotient of the standard uncertainty and the mean are of order unity also arise, e.g., in dimensional metrology and in photometry and radiometry. There are also problems in *limits of detection* (Section 9.5), where the uncertainties involved are comparable to the magnitudes of the quantities measured.

The effect of bias in the evaluated endpoints of the coverage interval constructed in this way, resulting from the use of a finite sample, can be reduced using so-called bias-corrected methods [30].[10]

### 7.3.1 Computation time

An indication of the computation time required for Monte Carlo calculations can be obtained from the following figures.

A problem with a model consisting of the sum of five terms, a cosine, a sine, an inverse tangent, an exponential and a square root was synthesised. Each term was assigned a Gaussian pdf. $M = 10^6$ Monte Carlo trials were made. Computation time for a 1 GHz Pentium 3 PC using Matlab are as follows.

The generation of the $5M$ Gaussian pseudo-random numbers takes 1 s.

The evaluation of the $M$ model values takes 1 s.

To sort the $M$ estimates of the output into non-decreasing order to produce the output distribution function takes 3 s.[11]

Thus, the computation time totals 5 s.

This information provides a simple basis for estimating the computation time for other models, other values of $M$ and other PCs.

## 7.4 Monte Carlo Simulation for multivariate models

Consider the counterpart of Section 7.2 for multivariate models. The model is now

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}).$$

[10] In the authors' experience these corrections typically have a small effect. This aspect, however, merits further study.

[11] The sorting should be carried out using a sorting algorithm that takes a number of operations proportional to $M \log M$ [55]. A naive sorting algorithm would take a number of operations proportional to $M^2$ that might make the computation time unacceptably long.

$M$ samples $\mathbf{x}_i$ of the input quantities $\mathbf{X}$ are taken as before. For each $\mathbf{x}_i$, evaluate the model as previously, except now the values of the output quantities are $\mathbf{y}_i = \mathbf{f}(\mathbf{x}_i)$, $m \times 1$ vectors.

Assemble these vector output quantities into an $m \times M$ matrix[12]:

$$\Upsilon = (\mathbf{y}_1, \ldots, \mathbf{y}_M).$$

From this matrix the covariance matrix $V_{\mathbf{y}}$ can be estimated. It is readily verified that

$$V_{\mathbf{y}} = \frac{1}{M-1} \Upsilon'(\Upsilon')^{\mathrm{T}},$$

where $\Upsilon'$ is $\Upsilon$ corrected for the sample means over all $M$ trials, i.e., with the mean of the $j$th row subtracted from all elements in that row, for $j = 1, \ldots, M$.

This covariance matrix contains (generally a more reliable estimate of) the information that would be delivered by a linear analysis such as Mainstream GUM. (In fact, it provides more than Mainstream GUM, since that procedure does not in general cover multivariate models.) The matrix $\Upsilon$ provides much richer information, however, in the following sense. Consider *any* function of the output quantities, i.e., the "next" output quantities in a subsequent stage of a multi-stage process for which the input quantities are those output quantities. The function evaluated for any column of $\Upsilon$ is an instance of the next output quantity. Since this output quantity can be calculated for all columns, the resulting $1 \times M$ row vector providing an empirical sampling distribution for that quantity. In particular, percentiles of the values in that vector can be used, as above, to provide a coverage interval for the derived quantity. Another quantity could be so introduced and the two row vectors used to compute any statistics required (mean, median, etc.) and the pair of vectors used to estimate correlation effects. Thus, the matrix $\Upsilon$ is a very valuable array, being the basis of almost unlimited statistical information.

The extension of the approach to the evaluation of coverage regions for multivariate measurement results is not straightforward, because the operation of sorting multivariate data is generally not well-defined. Some approaches have been proposed [3], including the ranking of multivariate data using the metric

$$(\mathbf{y}_i - \mathbf{a})^{\mathrm{T}} \Sigma^{-1} (\mathbf{y}_i - \mathbf{a}), \tag{7.1}$$

where $\mathbf{a}$ is a location statistic, such as the mean or (spatial) median [58], for the set $\mathbf{y}_i$ and $\Sigma$ is a dispersion statistic, such as the covariance matrix $V_{\mathbf{y}}$, for the set.

The provision of coverage regions in general is currently a research topic. A simple, practical approach is therefore proposed for current purposes. As indicated in Section 3.1, even in the univariate case the coverage interval is not unique. There is far greater freedom of choice in the multivariate case, where any domain containing 95% of the distribution of possible values constitutes a 95% coverage region. Moreover, a coverage interval can be expressed in terms of just two quantities, such as the interval endpoints or the interval midpoint and the semi-width. In more than one dimension, there is an infinite number of possibilities for the *shape* of the region.

[12]The symbol $\Upsilon$ is (reluctantly) used to denote the matrix of $y$-vectors, since $Y$ is used to denote a scalar output quantity and $\mathbf{Y}$ a vector output quantity.

A working approach is as follows. For linear or linearized problems the covariance matrix of the (vector) output quantity defines a one-standard-deviation ellipsoid [51] centred on the point denoting the estimate of the output quantity. Ellipsoids concentric with this one contain various fractions of the distribution of values attributed to the output quantity. For a given level of probability, 95%, say, the size of the ellipsoid from this set can be found (using the theory of multidimensional Gaussian distributions) that contains 95% of the possible values of the output. Such an ellipsoid can be constructed from the above covariance matrix, but its *size* would be dictated by the Gaussian assumption and not depend on the actual distribution of the values of $\mathbf{y}_i$. *An* ellipsoid is required that contains 95% of the actual distribution. In the univariate case, it is more valid, as considered in Section 7.3, to derive the coverage interval from the information contained in the empirical estimate of the actual pdf. Similarly, in the multivariate case the points $\mathbf{y}_i$ define a cluster of points centered on $\mathbf{y}$. These $\mathbf{y}_i$ can be expected to reflect faithfully the actual distribution, as a consequence of the sampling approach used, Therefore, it is recommended to define *a* coverage region by the (first-order) ellipsoid that (just) contains 95% of these $\mathbf{y}_i$.

It is emphasized that this approach will provide *a* 95% coverage region. The extent to which it is appropriate depends on the context. It may be highly inappropriate if the actual distribution of points $\mathbf{y}_i$ is, say, star-shaped. However, the approach is consistent with the use of the metric (7.1) with $\Sigma = V_{\mathbf{y}}$.

## 7.5 Extensions to implicit or complex-valued models

The extension of the above concepts to implicit models is conceptually straightforward. Instead of forming values $y_i = f(\mathbf{x}_i)$ of the output quantity $Y$ in the univariate case or $\mathbf{y}_i = f(\mathbf{x}_i)$ of $\mathbf{Y}$ in the multivariate case, by evaluating a formula or formulae, it is necessary to solve, respectively, the equation $h(y_i, \mathbf{x}_i) = 0$ for $y_i$ or the equations $\mathbf{h}(\mathbf{y}_i, \mathbf{x}_i) = \mathbf{0}$ for $\mathbf{y}_i$. (See Chapter 6.)

## 7.6 Disadvantages and advantages of MCS

There are several disadvantages of MCS and a (greater) number of advantages. The attributes of the approach are briefly reviewed.

### 7.6.1 Disadvantages

1. *Availability of pseudo-random number generators.* Pseudo-random number generators are required that are appropriate for the specified pdf's and joint pdf's of the input quantities that are likely to arise in metrology.

2. *Quality of pseudo-random number generators.* Some pseudo-random number generators are known to yield sequences of values that fail to satisfy standard tests for randomness.

3. *Repeatability properties.* The results may not be repeatable, thus making the testing of MCS software harder. The same random number generator,

using the same seed, must be employed to provide exactly repeatable results.

4. *Sensitivity coefficients.* Sensitivity coefficients are not immediately available. MCS does not automatically provide these coefficients. See, however, Appendix E.

5. *Complicated models.* The computational time required to carry out a sufficient number of MCS trials may be prohibitive if the model is complicated. See Section 7.8.

6. *Model evaluation.* In MCS the model is evaluated for each sample of the input quantities and hence for a range of values (that may be a number of "standard deviations" away from the estimates of the input quantities). This is in contrast to the Mainstream GUM procedure in which the measurement model is evaluated only at the estimates of the input quantities. For this reason some issues may arise regarding the numerical procedure used to evaluate the model, e.g., ensuring its convergence (where iterative schemes are used) and numerical stability.

### 7.6.2 Advantages

1. *Straightforward use.* Software can be implemented such that the user provides information concerning just the model and the parameters defining the pdf's of the input quantities.

2. *An estimate of the pdf of the output quantity (for univariate problems) is provided* (rather than a single statistic such as the standard deviation). Any required statistic (standard deviation, higher-order moments, etc.), coverage intervals and derived statistics such as the uncertainties of any function of the output $Y$ can be calculated from this distribution.

3. *An estimate of the (joint) output pdf for multivariate problems is provided.* This takes the form of a set of ($M$) values (points) in the space of the output quantities. This information is valuable in the context of multi-stage models in which the output from one stage becomes the input to the next. Sampling from these points embodies all the distributional information (including correlation) that is present.

4. *Applicability to a wide range of models.* MCS is broadly applicable regardless of the nature of the model:

   (a) The model may be linear, mildly nonlinear or strongly nonlinear. No initial analysis of the model is required to decide, for instance, how many terms in the Taylor-series expansion are required to approximate $f$ adequately for purposes of determining unbiased estimates of statistics associated with the measurement result.

   (b) The uncertainties of the input quantities may be arbitrarily large.

   (c) No assumption is made concerning the pdf of the output quantity $Y$. Thus, distributions that cannot be negative for instance, such as a distribution of distances, can be properly estimated.

5. *Symmetry is not assumed.* No assumption is made in using MCS concerning the symmetry of the input quantities or the output quantity. Thus, there is no need to "symmetrize" any pdf, or indeed any advantage gained from doing so.[13]

6. *Derivatives are not required.* There is no need for algebraic expressions for the first-order partial derivatives of the model with respect to the input quantities and for the evaluation of these expressions at the mean values of the input quantities.

7. *Avoidance of the concept of effective degrees of freedom.* MCS avoids the concept of effective numbers of degrees of freedom: an experimental mean and a standard deviation of a quantity, for which a Gaussian prior has been assumed, are described by a posterior density, viz., a linearly transformed Student's-$t$ distribution with the mean as the location parameter and the standard deviation as the scale parameter.[14]

8. *Linear computing time.* The computing time is dominated by the product of the number of trials and the time to evaluate the model $f$ for a set of input values. Over and above this, it is independent of the number $n$ of inputs. (This is not the case for the numerical evaluation of the multivariate integral (3.1) that defines $Y$, where the computation time is essentially proportional to $C^n$, for some constant $C$.)

9. *Sensitivities can be calculated.* MCS does not automatically provide sensitivity coefficients, for two reasons. First, they are not required for purposes of its operation. Second, for a nonlinear model, sensitivity coefficients are in general approximate, the quality of the approximations worsening with increased standard deviations for the input pdf's. However, simply by holding all input quantities but one fixed at their mean values MCS can be used to provide the pdf for the output quantity for the model having just that input quantity. See Section E.

10. *Multi-stage models.* MCS can take the output matrix $\Upsilon$ of vectors $\mathbf{y}_i$ from one stage of a multi-stage model, and carry out bootstrap-like re-sampling at the input to the next.

## 7.7 Implementation considerations

This section is concerned with two implementation aspects of MCS, (i) sampling from the pdf's of the input quantities and (ii) carrying out an appropriate number of Monte Carlo trials.

[13] The (UK) Royal Society of Chemistry state [17] that "Such an assumption [the use of a single parameter–often taken to be the half-range] is appropriate only if the "dispersion of values" is nearly symmetric about the measured result. It is easy to think of reasons why this may be false, including the effects of contamination and of range shifting on instruments".

[14] A discussion [36] of this issue suggests that in the case of a finite number of degrees of freedom the standard deviation $u$ of the Gaussian distribution should also be regarded as a random variable. In this regard, a pdf should also be attached to $u$. Because of the Gaussian assumption this pdf is distributed as $\chi^2$. A pseudo-code is available [36] that accounts for this effect.

Probability density functions that are commonly assigned to the input quantities are the uniform, the Gaussian, the Student's-$t$, and the multivariate Gaussian. Methods, in the form of pseudo-random number generators, for sampling from these pdf's are available in a companion document [20].

Other pdf's arise from time to time: Binomial, Gamma (including Exponential and $\chi^2$, Beta, Poisson, truncated Gaussian, etc.). The next edition of this document will cover such pdf's. Moreover, a general approach that is applicable to any pdf will be outlined.

Knowing when to stop the MCS process is discussed below. This aspect is important in terms of (a) carrying out an adequate number of trials in order that the process has "converged" according to a specified threshold, and (b) not carrying an excessive and hence uneconomic number of trials.

### 7.7.1 Knowing when to stop

When applying MCS the quality of the results depends on the number of MCS trials made. It is not possible to say *a priori* how many trials will be required to achieve a prescribed accuracy in the results. The reason is that the number will depend on the "shape" of the pdf and the level of probability required. Also, the calculations are *stochastic* in nature, being based on random sampling. Therefore, any discussion concerning "convergence" cannot be couched in the traditional numerical analysis setting of iterative methods for deterministic problems.

Here, a recommendation is made for determining the number of trials needed to provide a coverage interval at, for definiteness, a 95% level of probability to within a stipulated accuracy. It is necessary to define what is meant by accuracy. The following definition accords with the general sentiments of the GUM in reporting uncertainty. It is also consistent with earlier advice [10]. Suppose that two correct figures[15] are required in the standard deviation $u(y)$ of the measurement result $y$. The actual value of $u(y)$ will have a certain number $q$, say, of digits after the decimal point given by rounding that value to two significant decimal figures.

**Example 22** *The number of digits after the decimal point in rounding a standard uncertainty*

Suppose the measurement result for a nominally 100 g standard of mass [1, Clause 7.2.2] is $y = 100.021\ 47$ g and the standard uncertainty of $y$, rounded to two significant figures, is $u(y) = 0.000\ 35$ g, The number of digits in $u(y)$ after the decimal point is $q = 5$.

A coverage interval whose endpoints are known to a specified accuracy (viz., $q$ correct digits after the decimal point) is sought. However, because the sampling process is stochastic, no *guarantee* can be given that the required accuracy has been achieved. Instead the practical goal of determining the endpoints of the coverage interval to meet a *statistical* criterion is set. Within the context of the sampling process, it is possible to state a 95% coverage interval for the *endpoints* of a computed coverage interval.

[15] The approach is easily modified for other numbers of correct figures. Two figures (or often one figure) are nearly always adequate.

It was stated in Section 7.3 that after $M$ MCS trials, the endpoints of a coverage interval for $Y$ are the 2.5-and 97.5-percentiles $y_{(0.025M)}$ and $y_{(0.975M)}$. The endpoints of a coverage interval for the $(100p)$th percentile are [6] $Mp \pm 2\{Mp(1-p)\}^{1/2}$. For the 2.5- and 97.5-percentiles, these endpoints are $0.025M \pm 0.31M^{1/2}$ and $0.975M \pm 0.31M^{1/2}$. Let

$$\begin{aligned} L- &= [0.025M - 0.31M^{1/2}], \\ L &= [0.025M], \\ L+ &= [0.025M + 0.31M^{1/2}], \\ H- &= [0.975M - 0.31M^{1/2}], \\ H &= [0.975M], \\ H+ &= [0.975M + 0.31M^{1/2}]. \end{aligned}$$

$M$ is large enough when the lengths $y_{(L+)} - y_{(L-)}$ and $y_{(H+)} - y_{(H-)}$ of the 95% coverage intervals for the endpoints $y_{(L)}$ and $y_{(H)}$ are no larger than $0.5 \times 10^{-q}$,

A practical and easily implemented approach consists of carrying out $M = 5,000$ (say) trials initially, calculating the value of $u(y)$ from these results, forming the value of $q$, computing the coverage interval and carrying out the above accuracy check. If the check was satisfied, accept the result. Otherwise, carry out a further 5000 trials, and recompute $u(y)$, $q$ and the coverage interval based on the aggregated (increased) number of trials and carry out the checks again. Repeat the process as many times as necessary.[16]

**Example 23** *An appropriate number of Monte Carlo trials to be taken when determining the measurement uncertainty of a hand-held multimeter*

Consider the example of Section 9.3 concerned with the calibration of a hand-held multimeter. Table 7.1 shows the application of the above procedure. The first column in the table gives the total number of trials taken. The second column gives the number of digits after the decimal point in the value of $u(y)$, the standard deviation of the $M$ values $y_i$ of the measurand $Y \equiv E_X$ determined so far. The third column contains the value of $u(y)$. The columns headed $y_{(L)}$ and $y_{(H)}$ give the endpoints of the coverage interval formed from the appropriate percentiles of the $y_i$. The columns headed $y_{(L-)}$ and $y_{(L+)}$ give the lower and upper endpoints of the coverage interval for $y_{(L)}$, and $y_{(H-)}$ and $y_{(H+)}$ similarly for $y_{(H)}$. The columns headed $L$-len and $H$-len give the lengths of these endpoint coverage intervals.

It is seen that these lengths reduce approximately as $M^{-1/2}$, in accordance with the point made in Section 7 concerning convergence.[17] 60,000 trials were needed to reduce both of these lengths to 0.0005, and thus to give a good degree of assurance that the coverage interval for $Y \equiv E_X$ was computed correctly to three figures after the decimal point, in accordance with providing two significant figures in $u(y)$. The resulting coverage interval for the measurand, defined by

[16] Because the value of $u(y)$ is much better defined by a relatively small number of trials than are the endpoints of a coverage interval, subsequent values of $u(y)$ are unlikely to change much from their original values. In contrast, the coverage interval will take longer to settle down.

[17] For instance, when $M$ is increased from 15 000 to 60 000, i.e., increased by a factor of four, $L$-len and $H$-len are approximately halved, reducing from 0.0009 to 0.0005 and from 0.0010 to 0.0005, respectively.

| M | $q$ | $u(y)$ | $y_{(L-)}$ | $y_{(L)}$ | $y_{(L+)}$ | $y_{(H-)}$ | $y_{(H)}$ | $y_{(H+)}$ | $L$-len | $H$-len |
|---|---|---|---|---|---|---|---|---|---|---|
| 5000 | 3 | 0.0296 | 0.0477 | 0.0487 | 0.0501 | 0.1498 | 0.1508 | 0.1517 | 0.0024 | 0.0019 |
| 10000 | 3 | 0.0295 | 0.0490 | 0.0495 | 0.0503 | 0.1502 | 0.1507 | 0.1514 | 0.0013 | 0.0012 |
| 15000 | 3 | 0.0297 | 0.0486 | 0.0491 | 0.0496 | 0.1501 | 0.1507 | 0.1512 | 0.0009 | 0.0010 |
| 20000 | 3 | 0.0295 | 0.0489 | 0.0494 | 0.0499 | 0.1501 | 0.1506 | 0.1511 | 0.0011 | 0.0010 |
| 25000 | 3 | 0.0296 | 0.0493 | 0.0498 | 0.0502 | 0.1503 | 0.1507 | 0.1511 | 0.0009 | 0.0008 |
| 30000 | 3 | 0.0296 | 0.0492 | 0.0496 | 0.0499 | 0.1504 | 0.1507 | 0.1510 | 0.0007 | 0.0007 |
| 35000 | 3 | 0.0296 | 0.0489 | 0.0493 | 0.0497 | 0.1501 | 0.1504 | 0.1508 | 0.0008 | 0.0007 |
| 40000 | 3 | 0.0296 | 0.0493 | 0.0496 | 0.0499 | 0.1503 | 0.1506 | 0.1510 | 0.0006 | 0.0007 |
| 45000 | 3 | 0.0296 | 0.0492 | 0.0495 | 0.0497 | 0.1502 | 0.1505 | 0.1508 | 0.0005 | 0.0006 |
| 50000 | 3 | 0.0296 | 0.0490 | 0.0493 | 0.0495 | 0.1503 | 0.1506 | 0.1509 | 0.0006 | 0.0006 |
| 55000 | 3 | 0.0296 | 0.0492 | 0.0495 | 0.0498 | 0.1503 | 0.1505 | 0.1508 | 0.0005 | 0.0005 |
| 60000 | 3 | 0.0296 | 0.0491 | 0.0494 | 0.0496 | 0.1502 | 0.1504 | 0.1506 | 0.0005 | 0.0005 |

Table 7.1: An illustration of the "convergence" of the MCS procedure, when applied to the model of a digital multimeter of Section 9.3. The first column in the table gives the total number of trials taken. The second column gives the number of digits after the decimal point in the value of $u(y)$, the standard deviation of the $M$ values $y_i$ of the measurand $Y \equiv E_X$ determined so far. The third column contains the value of $u(y)$. The columns headed $y_{(L)}$ and $y_{(H)}$ give the endpoints of the coverage interval formed from the appropriate percentiles of the $y_i$. The columns headed $y_{(L-)}$ and $y_{(L+)}$ give the lower and upper endpoints of the coverage interval for $y_{(L)}$, and $y_{(H-)}$ and $y_{(H+)}$ similarly for $y_{(H)}$. The columns headed $L$-len and $H$-len give the lengths of these endpoint coverage intervals.

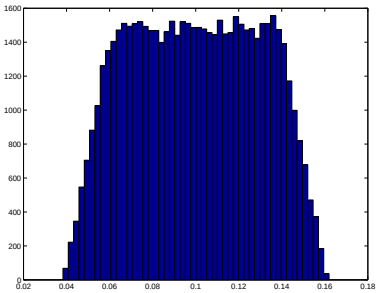


Figure 7.5: A histogram estimating the probability density function of the output of the digital voltmeter model produced using MCS.

the fifth and eighth values in the final row of the table, is [0.049, 0.150] V. This interval is to be compared with that in [28] of $(0.10 \pm 0.05)$ V or, equivalently, [0.05, 0.15] V, viz., agreement to the published number of figures.

Figure 7.5 shows a histogram that estimates the probability density function of the output of the digital voltmeter model. The histogram was produced using the final set of $M = 60,000$ model values $y_i$. It is essentially trapezoidal in shape, to be compared with the statement [28], made following an approximate analytical treatment, that the distribution is essentially rectangular.

Figure 7.6 shows the corresponding distribution function, estimated from the ordered $y_i$ values, as described in Section 7.3. The endpoints of the estimated 95% coverage interval are indicated in this figure by vertical lines.

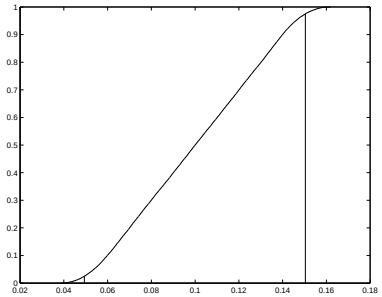


Figure 7.6: The distribution function of the output of the digital voltmeter model estimated using MCS. The endpoints of the estimated 95% coverage interval are indicated in this figure by vertical lines.

## 7.8 Summary remarks on Monte Carlo Simulation

MCS is a tool that is consistent with general GUM philosophy (GUM Clause G.1.5) and also with its interpretation [61] for scientists at the National Institute for Standards and Technology (NIST) in the United States. The major difference is that rather than propagating uncertainties through a linearized model, the pdf's of the input quantities are propagated through the model *per se* to calculate the pdf of the output quantity. From the pdf of the output quantity a coverage interval is obtained without making a Gaussian or any other assumption concerning the form of this pdf.

MCS can straightforwardly be applied to a range of uncertainty evaluation problems. For the most general such problem, it is emphasized that it would be necessary to provide

1. Pseudo-random number generators for the univariate and joint pdf's needed in the application.

2. A mechanism for determining coverage regions for multivariate results of measurement.

Recommendations ([20], Section 7.7.1) are intended to assist in this regard.

The degree of belief in the pdf's of the input quantities can be considered by repeating a simulation after having varied these functions. The sensitivity of the results to such critical information can thus be investigated.

For sinple models the number of Monte Carlo trials can be chosen to be substantially large, e.g., $10^6$ (Section 7.3.1). A complete uncertainty calculation would take some five seconds on a 1 GHz Pentium PC. More than half this time is taken with *sorting* the Monte Carlo estimates of the measurement result.

For models of modest complexity, taking say 100 times longer to evaluate, to achieve a comparable quality of result would take of the order of 200 seconds. The figure is not 500 seconds because the sorting time remains about 3 seconds. Nevertheless the computation time is now noticeable, particularly if many similar calculations have to be carried out. In such cases it would be desirable to consider the use of an automatic stopping rule (Section 7.7.1), rather than fixing the number of Monte Carlo trials in advance.

For very complicated models[18] it would not be economic to take more than a small number of trials (say 10). In such a case it would be impossible to provide a coverage interval reliably. Rather, a mean and standard deviation should be obtained and a coverage interval obtained assuming a Gaussian distribution. (Appeal can be made to the Principle of Maximum Entropy.) For multivariate output quantities a covariance matrix from the results can be calculated and a multivariate Gaussian distribution assumed. As always, the results obtained should be accompanied by a statement that indicates clearly how they were obtained.

[18] An instance is a model defined by a partial differential equation.

# Chapter 8

# Validation of Mainstream GUM

This chapter describes validation tests that can be carried out to decide whether Mainstream GUM is applicable in any particular circumstance. There are two types of test, prior tests and post tests, carried out, respectively, before utilizing Mainstream GUM and after having applied Mainstream GUM.

The post-test involves the use of MCS as the validation tool.

## 8.1 Prior test: a quick test of whether model linearization is valid

Mainstream GUM uses the first-order Taylor expansion of the model as part of evaluating the uncertainty associated with the result of measurement.

A quick test is described that can be used to help decide whether the first-order Taylor series approximation employed by Mainstream GUM is adequate. The test applies to a model with uncorrelated input quantities. Using the notation of Section 6.2.1, consider the standard uncertainty $u(y)$ in the output quantity $y$ obtained from

$$u^2(y) = \sum_{i=1}^{n} c_i^2 u^2(x_i),$$

a formula obtained from (6.2) when all covariances are set to zero, and $c_i$ denotes $\partial f/\partial X_i$ evaluated at $\mathbf{X} = \mathbf{x}$. The test is based on generating, for each model input $X_i$, not just a single value of the sensitivity coefficient $c_i$ but three values that are realistically likely to encompass the range of values of the first-order partial derivative for that input quantity in the neighbourhood of its estimated value. If the variation in these values is small enough, for all input quantities, a first-order evaluation is adequate. "Small enough" means that the variation is no greater than the required numerical accuracy of the solution uncertainty. "In the neighbourhood" means, for each $i$, the interval about $x_i$ that covers, say, 95% of the distribution of values of the input quantity. Thus, if one correct figure in a relative sense is sought, this relative variation is not to exceed 0.05. Compare with Section 7.7.1.

This test is severe. If just one of the model input quantities failed the test, the test overall would be regarded as failing, which would be unreasonable if that quantity made a negligible contribution to $u(y)$.

A less extreme and more practical test is (i) to evaluate $u(y)$ for the smallest of the three sensitivity coefficients for each component, (ii) as (i) but for the largest sensitivity coefficient, and (iii) to test whether the two estimates of $u(y)$ agreed to one figure.

For each value of $i$ from 1 to $n$, take the variation over the range $x_i^{(\mathrm{low})}$ to $x_i^{(\mathrm{high})}$, the 2.5- and 97.5-percentiles of the pdf of $x_i$.[1] The rationale for this choice is that the pdf of $y$ is likely to be dominated by contributions from the input pdf's that are "sufficiently central" to these pdf's: taking the stated values reflects this consideration. The variation of $c_i$ regarded as a function of $X_i$, i.e., $c_i = c_i(X_i)$, is thus considered over the interval $x_i \pm k_i u(x_i)$, where $k_i$ is the coverage factor such that 95% of the pdf is embraced.[2] It is therefore given by the interval $[c_i^{\min},\ c_i^{\max}]$, where

$$c_i^{\min} = \min c_i(X_i), \qquad c_i^{\max} = \max c_i(X_i),$$

where the minimum and maximum are taken over

$$X_i \in [x_i - k_i u(x_i),\ x_i + k_i u(x_i)].$$

It is considered unnecessary to determine the *actual* minimum and maximum values in this interval. In the great bulk of cases $c_i(X_i)$ would change monotonically over the interval, in which case the values at the endpoints would deliver the required values. If the interval contains an extreme value strictly within it, it is more reliable to take

$$c_i^{\min} = \min C_i, \qquad c_i^{\max} = \max C_i,$$

where $C_i$ denotes the set of three values

$$\{c_i(x_i - k_i u(x_i)),\ c_i(x_i),\ c_i(x_i + k_i u(x_i))\}.$$

This variant is the one recommended[3], particularly because the third piece of information, $c_i(x_i)$, would have been evaluated in any case as part of the conventional evaluation of $u(y)$.

The principle can be implemented in the following way.

1. For each value $i = 1, \ldots, n$

[1] For an input quantity $X_i$ with a Gaussian pdf, $x_i \pm 1.96u(x_i)$ can be taken. For an input quantity $X_i$ with a uniform pdf, take $x_i \pm 1.65u(x_i)$. The multiplier 1.65 is such that 95% of the uniform pdf is covered. In general, take the appropriate 95% coverage interval.

[2] This interval applies to a symmetric pdf. If the pdf is asymmetric the interval becomes $[x_i - k_i^{(-)}u(x_i),\ x_i + k_i^{(+)}u(x_i)]$, where $k_i^{(-)}$ and $k_i^{(+)}$ are factors such that the interval is a 95% coverage interval for $X_i$. The modifications to be made to the subsequent considerations are straightforward.

[3] It is possible to carry out more sophisticated calculations. One approach is to pass a quadratic polynomial through the three sensitivity values, and take its minimum or maximum as appropriate, if such a value lies within the interval. Another possibility is to evaluate $c_i(X_i)$ on a fine mesh of values in the interval in order to determine extreme values. Such strategies could be adopted, but their use would be inconsistent with the intent of the "quick test" of this section.

(a) Set $k_i$ such that $[x_i - k_i u(x_i),\ x_i + k_i u(x_i)]$ is a 95% coverage interval for $X_i$.

(b) Form $c_i^{(\mathrm{mid})} = \partial f(\mathbf{X})/\partial X_i$, evaluated at $\mathbf{x}$.

(c) Replace the $i$th component of $\mathbf{x}$ by $x_i^{(\mathrm{low})} = x_i - k_i u(x_i)$.

(d) Form $c_i(\mathrm{low}) = \partial f(\mathbf{X})/\partial X_i$, evaluated at $\mathbf{x}$.

(e) Replace the $i$th component of $\mathbf{x}$ by $x_i^{(\mathrm{high})}. = x_i + k_i u(x_i)$.

(f) Form $c_i^{(\mathrm{high})} = \partial f(\mathbf{X})/\partial X_i$, evaluated at $\mathbf{x}$.

(g) Restore the original value of $\mathbf{x}$.

(h) Form $c_i^{(\mathrm{min})} = \min(c_i^{(\mathrm{low})}, c_i^{(\mathrm{mid})}, c_i^{(\mathrm{high})})$.

(i) Form $c_i^{(\mathrm{max})} = \max(c_i^{(\mathrm{low})}, c_i^{(\mathrm{mid})}, c_i^{(\mathrm{high})})$.

2. Form $u^{(\mathrm{min})}(y)$ and $u^{(\mathrm{max})}(y)$ from

$$\left(u^{(\mathrm{min})}(y)\right)^2 = \sum_{i=1}^{n} \left(c_i^{(\mathrm{min})}\right)^2 u^2(x_i)$$

and

$$\left(u^{(\mathrm{max})}(y)\right)^2 = \sum_{i=1}^{n} \left(c_i^{(\mathrm{max})}\right)^2 u^2(x_i).$$

3. The prior test is satisfied if

$$u(y) - u^{(\mathrm{min})}(y) \leq 0.05u(y)$$

and

$$u^{(\mathrm{max})}(y) - u(y) \leq 0.05u(y).$$

If two correct figures are required, the value of 0.05 in the above procedure is to be replaced by 0.005.

## 8.2 Validation of Mainstream GUM using MCS

Any case of doubt in the applicability of Mainstream GUM should be validated. In an instance where the Mainstream GUM approach was indeed validated, practitioners could safely continue to employ it in that circumstance or in situations that could be regarded as comparable. If it was not so validated, it would be necessary to use an alternative procedure in future. It is recommended that the process (MCS) at the core of the validation procedure is used for that purpose.

It is essential to note that even if Mainstream GUM had been validated in terms of the coverage interval at a specific level of probability, it does not follow that the validation extends to coverage intervals at other levels of probability. For instance, many pdf's differ in the manner in which their "tails" behave. Thus, good agreement in the 95% coverage intervals that have been provided by Mainstream GUM and MCS by no means applies to, say a 99% level. This point is reinforced by the fact that the graphs of a number of pdf's tend to "intersect" not far from the endpoints of a 95% coverage interval.

A validation procedure can be based on the following points.

1. Given: Mainstream GUM results in the form of
   (a) a measurement result $y$,
   (b) a combined standard uncertainty $u(y)$,
   (c) a 95% coverage interval $y \pm U$, where $U$ denotes expanded uncertainty determined as part of the Mainstream GUM procedure.
2. Given: MCS results in the form of
   (a) a measurement result $y^{(\mathrm{MCS})}$,
   (b) a corresponding standard uncertainty $u(y^{(\mathrm{MCS})})$,
   (c) the endpoints $y_{\mathrm{low}}^{(\mathrm{MCS})}$ and $y_{\mathrm{high}}^{(\mathrm{MCS})}$ of a 95% coverage interval for the measurand.
3. Define $q$ to be the number of figures after the decimal point in $u(y^{(\mathrm{MCS})})$, after this value has been rounded to two, say, significant decimal digits.
4. Aim: determine whether there are $q$ correct digits after the decimal point in $u(y)$ and the endpoints $y - U$ and $y + U$.
5. Approach: compare the Mainstream GUM and MCS results to determine whether the required number of correct digits has been obtained.

Further details are available [20].

# Chapter 9

# Examples

The examples in this chapter are intended to illustrate the principles contained in the body of this guide. Where appropriate, two approaches, Mainstream GUM and Monte Carlo Simulation, are used and contrasted. Some examples can be regarded as typical of those that arise in metrology. Others are more extreme, in an attempt to indicate the considerations that are necessary when those of "normal circumstances" fail to apply. Perhaps, unfortunately, such adverse circumstances arise more frequently than would be wished, in areas such as limit of detection, electromagnetic compliance, photometry and dimensional metrology.

Many other examples are given throughout this guide, some to illustrate basic points and others more comprehensive.

## 9.1 Flow in a channel

This example was provided by the National Engineering Laboratory. It concerns an implicit model arising in channel flow.

Open channel flows are common in the water and hydroelectric power industries and where river extraction provides cooling water for industrial processes. Such a flow can be measured by introducing a specially constructed restriction in the flow channel. The flow is then a function of the geometry of the restriction (width upstream, throat width and length, height of the hump in the floor of the restriction) and the depth of water passing through the restriction.

The model input quantities and estimates of their values are:

| | | | |
|---|---|---|---|
| Approach channel width | $B = 2.0$ m, | Throat width | $b = 1.0$ m, |
| Hump height | $p = 0.25$ m, | Throat length | $L = 3.0$ m, |
| Nominal head | $h = 1.0$ m | | |

The output quantity is the flow rate $Q$. The model relating $Q$ to the input quantities is

$$Q = (2/3)^{3/2} g^{1/2} C_v C_D b h^{3/2},$$

with $g = 9.812$ m$s^{-2}$, the acceleration due to gravity,

$$C_D = (1 - 0.006L/b)(1 - 0.003L/h)^{3/2} \tag{9.1}$$

and

$$4b^2h^2C_v^2 - 27B^2(h+p)^2(C_v^{2/3} - 1) = 0. \tag{9.2}$$

To calculate $Q$ for any values of the input quantities, it is first necessary to form $C_D$ from the explicit formula (9.1) and $C_v$ from the implicit equation (9.2).[1]

The first four input quantities are geometric dimensions obtained by a series of measurements with a steel rule at various locations across the flume. There are uncertainties in these measurements due to location, rule reading errors and rule calibration. Head height is measured with an ultrasonic detector, uncertainties arising from fluctuations in the water surface and instrument calibration.

All uncertainty sources were quantified and appropriate pdf's assigned. All pdf's were based on Gaussian or uniform distributions. The standard deviations of these pdf's (standard uncertainties of the input quantities) were all less than 0.3% relative to the means (estimated values).

Both the Mainstream GUM procedure (Section 6) and the Monte Carlo procedure (Section 7) were applied. The Mainstream GUM results were validated (Section 8) using the Monte Carlo procedure, under the requirement that results to two significant figures were required. In fact, the coverage interval for $y$ as produced by Mainstream GUM was confirmed correct (by carrying out further Monte Carlo trials) to *three* significant figures. To quote the comparison in a relative sense, the quotient of (a) the half-length of the 95% coverage interval for $y$ ($\equiv Q$) and (b) the standard uncertainty of $y$ was 1.96, which agrees to three significant figures with the Mainstream GUM value, viz., the (Gaussian) coverage factor for 95% coverage. For further comparison, the corresponding quotients corresponding to 92.5% and 97.5% coverage intervals were 1.78 and 2.24, also in three-figure agreement with Mainstream GUM results. It is concluded that the use of Mainstream GUM is validated for this example for the coverage probabilities indicated.

## 9.2 Graded resistors

This example is intended to cover an instance where it would be important to take specific account of the pdf of an input quantity to help ensure that valid results are obtained from an uncertainty evaluation.

The uncertainties of a mass-produced electrical circuit are to be evaluated. The circuit contains electrical components of various types. One of these component types, a resistor, is considered here.

High-grade (nominally) 1 Ω resistors are graded according to their specification. A-grade resistors are those that lie within 1% of nominal, B-grade within 5% and C-grade within 10%. The allocation of resistors to the various grades is decided by measurement. For the purposes of this example, the uncertainty of this measurement is taken as negligible. As each resistor is measured it is allocated to an A-grade, a B-grade, a C-grade or an unclassified "bin". The allocation is made in the following sequential manner. If a resistor has resistance in the interval $(1.00 \pm 0.01)$ Ω it is allocated to the A-grade bin. If not, and it has resistance in the interval $(1.00 \pm 0.05)$ Ω it is allocated to the B-grade bin.

[1] This equation is in fact a cubic equation in the variable $C_v^{2/3}$ and, as a consequence, $C_v$ *can* be expressed explicitly in terms of $B$, $h$ and $p$. Doing so is unwise because of the possible numerical instability due to subtractive cancellation in the resulting form. Rather, the cubic equation can be solved using a recognised stable numerical method.

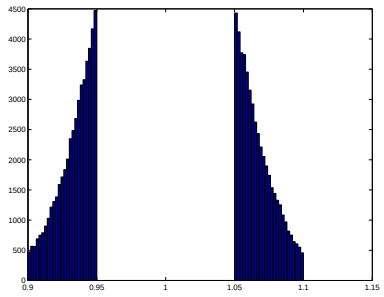


Figure 9.1: A histogram produced by MCS to estimate the probability density function for the Grade-C resistor.

If not, and it has resistance in the interval $(1.00 \pm 0.10)\ \Omega$ it is allocated to the C-grade bin. Otherwise, it is allocated to the unclassified bin.

For the circuit application, C-grade resistors are selected. All such resistors have a resistance in the interval $[0.90,\ 0.95]\ \Omega$ or the interval $[1.05,\ 1.10]\ \Omega$. From the knowledge of the manufacturing process, the pdf of the resistance of a resistor before the allocation process is carried out can be taken as Gaussian with mean $1.00\ \Omega$ and standard deviation $0.04\ \Omega$.

Consider the use of three such resistors in series within the circuit to form a (nominally) $3\ \Omega$ resistance.

The model for the $3\ \Omega$ resistance $R$ is

$$R = R_1 + R_2 + R_3,$$

where $R_i$ denotes the resistance of resistor $i$. Each $R_i$ has a pdf as above. What is the pdf of $R$ and what is a 95% coverage interval for $R$?

Figures are used to show diagrammatically histograms and distribution functions of the results of using MCS to estimate the pdf's. An analytic or semi-analytic treatment is possible, but MCS enables results to be provided rapidly. All results are based on the use of $M = 10^5$ Monte Carlo trials.

Figure 9.1 shows a histogram of the pdf of $R_i$. It is basically Gaussian, with the central and tail regions removed as a consequence of the grading process.

Figure 9.2 shows the corresponding distribution function produced by MCS in the manner of Section 7.3 The endpoints of the estimated 95% coverage interval are indicated by vertical dashed lines and the corresponding endpoints under the Gaussian assumption by vertical dotted lines.

Figure 9.3 shows a histogram produced by MCS to estimate the probability density function for $R$, three Grade-C resistors in series.

The pdf is multimodal, possessing four maxima. The mean of the pdf is $3.00\ \Omega$, the sum of the means of the three individual pdf's, as expected. This value is, however, unrepresentative, an "expectation" that could rarely occur.

Figure 9.4 shows the corresponding distribution function produced in the manner of Section 7.3.

The pdf of Figure 9.3 could be perceived as an overall bell-shape, with strong structure within it. Indeed, the counterpart of these results for six resistors in series, as illustrated in Figures 9.5 and 9.6, lies even more in that direction.

Table 9.1 summarises the numerical results, and also includes the results for $n = 10$ and 20 resistors. It is reassuring that, considering the appreciable

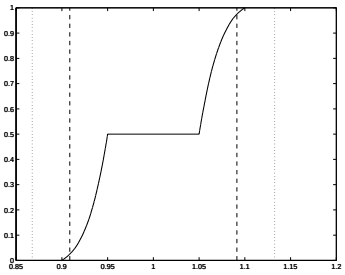

Figure 9.2: The distribution function corresponding to the probability distribution function of Figure 9.1. The endpoints of the estimated 95% coverage interval are indicated by vertical dashed lines and the corresponding endpoints under the Gaussian assumption by vertical dotted lines.

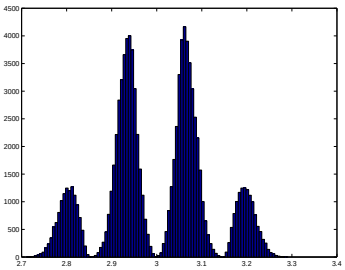

Figure 9.3: A histogram produced by Monte Carlo Simulation to estimate the probability density function for $R$, three Grade-C resistors in series.

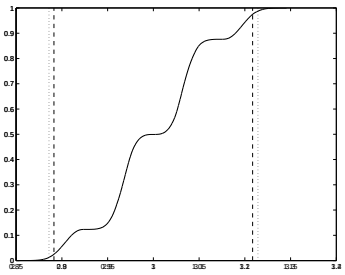

Figure 9.4: The distribution function corresponding to the probability distribution function of Figure 9.3. The endpoints of the estimated 95% coverage interval are indicated by vertical dashed lines and the corresponding endpoints under the Gaussian assumption by vertical dotted lines.

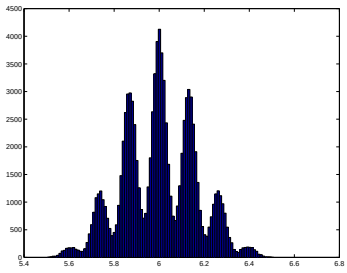


Figure 9.5: A histogram produced by Monte Carlo Simulation to estimate the probability density function for six Grade-C resistors in series.

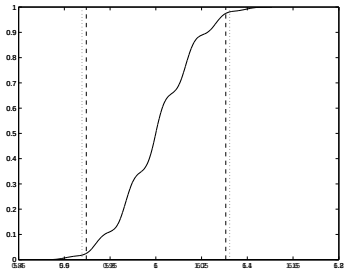


Figure 9.6: The distribution function corresponding to the probability distribution function of Figure 9.5. The endpoints of the estimated 95% coverage interval re indicated by vertical dashed lines and the corresponding endpoints under the Gaussian assumption by vertical dotted lines.

| No. $n$ of resistors | Standard uncertainty | Endpoints of 95% coverage intervals Monte Carlo | | Mainstream GUM | |
|---|---|---|---|---|---|
| 1 | 0.07 | 0.91 | 1.09 | 0.87 | 1.13 |
| 3 | 0.12 | 2.78 | 3.22 | 2.77 | 3.23 |
| 6 | 0.17 | 5.69 | 6.31 | 5.68 | 6.32 |
| 10 | 0.21 | 9.58 | 10.42 | 9.58 | 10.42 |
| 20 | 0.30 | 19.41 | 20.59 | 19.41 | 20.59 |

Table 9.1: Evaluation of 95% coverage intervals for $n$ Grade-C 1 Ω resistors in series using Monte Carlo Simulation and Mainstream GUM. The units of Columns 2-6 are ohms.

departure from normality, the coverage interval "converges" rapidly to that obtained under the assumption that the output pdf is Gaussian (as in Mainstream GUM). These are no grounds for complacency, however: there will be situations where the use of Mainstream GUM is not so favourable.

As stated, an analytical treatment would be possible for this problem. It might be difficult to justify the effort required, however, unless the analysis provided some general insight that would give added value to the application. Using existing MCS software, it required approximately one hour to enter the problem and produce the numerical and graphical results. The computation time itself was negligible, being a few seconds in all.

## 9.3 Calibration of a hand-held digital multimeter at 100 V DC

A hand-held digital multimeter (DMM) is calibrated at an input of 100 V DC using a multifunction calibrator as a working standard. A model for the error of indication $E_X$ of the DMM [28] is

$$E_X = V_{iX} - V_S + \delta V_{iX} - \delta V_S,$$

where the model input quantities and their pdf's are defined and assigned as follows:

**DMM reading $V_{iX}$.** The voltage indicated by the DMM (the index $i$ meaning "indication"). Because of the limited resolution of the device, no scatter is observed in the indicated values. Therefore, the indicated voltage, 100.1 V, at the calibrator setting of 100 V, is taken as exact.

**Voltage $V_S$ generated by the calibrator.** The calibration certificate for the calibrator states that the voltage generated is the value indicated by the calibrator setting and that the associated expanded uncertainty of measurement associated with the 100 V setting is $U = 0.002\ V$ with a coverage factor of $k = 2$. In the absence of other knowledge a Gaussian pdf with mean 100 V and standard deviation 0.001 V (obtained from $U/k = 0.002/2$) is therefore assigned to this input.

**Correction $\delta V_{iX}$ of the indicated voltage of the DMM.** The least significant digit of the DMM display corresponds to 0.1 V as a consequence of

the finite resolution of the instrument. Each DMM reading is therefore taken to have a correction of 0.0 V, the error in this value lying in the interval $\pm 0.05\ V$. In the absence of other knowledge, a uniform pdf with mean 0.0 V and standard deviation 0.029 V (obtained from $0.05/\sqrt{3}$) is therefore assigned to this input.

**Correction $\delta V_S$ of the calibrator voltage.** The calibrator voltage is in principle corrected for a range of effects including drift, mains power deviations and loading. An analysis [28] gives a correction of 0.0 V and states that the error in this value lies in the interval $\pm 0.011\ V$.[2] In the absence of other knowledge, a uniform pdf with mean 0.0 V and standard deviation 0.0064 V (obtained from $0.011/\sqrt{3}$) is therefore assigned to this input.

This model was analyzed using Monte Carlo Simulation. Details are given in Section 7.7.1. The error of indication of the DMM was found to be 0.100 V with a 95% coverage interval of [0.049, 0.150] V. The corresponding result obtained by an approximate analytical treatment [28] was $(0.10 \pm 0.05)$ V, i.e., in agreement to the figures quoted.

## 9.4 Sides of a right-angled triangle

This example is intended to illustrate the use of statistical modelling, as described in Section 4.2. It also illustrates the manner in which correlation effects can sometimes be removed by the introduction of an additional variable, as indicated in Section 4.2.1.

The sides of a right-angled triangle are repeatedly measured with a length-measuring instrument. The measurements contain random and systematic effects. Use all these measurements to determine best estimates of the sides of the triangle and evaluate their uncertainties.

Denote the shorter sides of the triangle by $A$ and $B$ and the hypotenuse by $H$. Let there be $n_A$ measurements of $A$, $n_B$ of $B$ and $n_H$ of $H$. Let the $i$th measurement of $A$ be $a_i$, with error $\Delta a_i$, and similarly for $B$ and $H$. The relationships between the measurements, the "true" values and the errors are

$$\begin{aligned} a_i &= A + \Delta a_i, \qquad i = 1, \ldots, n_A, \\ b_i &= B + \Delta b_i, \qquad i = 1, \ldots, n_B, \\ h_i &= H + \Delta h_i, \qquad i = 1, \ldots, n_H. \end{aligned}$$

According to Pythagoras' theorem, the sides are *physically* related by

$$A^2 + B^2 = H^2. \tag{9.3}$$

For consistency, the solution values of $A$, $B$ and $H$ are to satisfy this condition.

The errors $\Delta a_i$, etc. are *statistically* related in that the instrumental systematic effect will manifest itself in all these values. Its "presence" means that all measurements are correlated. In order to quantify this correlation, it is

[2] Since this correction is based on a number of effects, it does not seem reasonable (cf. Section 4.3.3 to regard the error as equally likely to take any value in this interval. However, since the effect of this input on the model output is relatively small, and since an intention is to compare the EA approach with that of Mainstream GUM, M3003 and MCS, the above form is taken.

conventionally necessary to know the standard deviation $u(\Delta L)$ of the instrumental systematic effect and the repeatability standard deviation $u_{\rm rep}$ of the measurements.

A covariance matrix based on this correlation can then be established[3] and the above equations solved by least squares with an "input covariance matrix" equal to this covariance matrix. Generic details of this approach, *Gauss-Markov estimation*, are available [22]. Formally, the result is in the form of a GUM model

$$\mathbf{Y} \equiv (A,\ B)^{\rm T} = \mathbf{f}(a_1, \ldots, a_{n_A}, b_1, \ldots, b_{n_B}, h_1, \ldots, h_{n_H}).$$

The measurements $a_i$, $b_i$ and $h_i$ are the input quantities ($n_A + n_B + n_H$ in number) and $A$ and $B$ are the (two) output quantities. The third side, $H$, the hypotenuse of the triangle, is not included as an output quantity, since it can be formed from $A$ and $B$ using (9.3). $\mathbf{f}$ denotes the model. It cannot, at least conveniently, be written down mathematically, but is defined by the computational procedure that implements the least-squares solution process.

Propagation of the input covariance matrix through the model to provide the covariance matrix of $(A,\ B)^{\rm T}$ can be carried out as discussed in Section 6.2. The use of Equation (9.3) as a "next-stage" model (cf. Section 4.2.3), providing the output $H$ in terms of $A$ and $B$, can then be used to determine the uncertainty of $H$. The results can be combined to provide the covariance matrix of $(A,\ B,\ H)^{\rm T}$.

Statistical modelling can be used to provide the required sides and their uncertainties without having to work with correlated quantities and, in this instance, without prior knowledge of the above standard deviations. Regard this systematic effect as an unknown bias $\Delta L$ and write

$$\Delta a_i = \Delta L + \delta a_i,$$

etc., where $\delta a_i$ is the random error in $a_i$, etc. The above relationships become

$$\begin{array}{rcll} a_i & = & A + \Delta L + \delta a_i, & i = 1, \ldots, n_A, \\ b_i & = & B + \Delta L + \delta b_i, & i = 1, \ldots, n_B, \\ h_i & = & H + \Delta L + \delta h_i, & i = 1, \ldots, n_H. \end{array}$$

The errors $\delta a_i$, etc. are uncorrelated, the covariance matrix being diagonal with all entries equal to $u^2_{\rm rep}$. Best estimates of the sides (and $\Delta L$) are then given by an ordinary least-squares problem (Gauss estimation). First, it is necessary to incorporate the condition (9.3) [22]. There are various ways to do so in general, but here it is simplest to use the condition to eliminate a variable. One possibility is to replace $H$ by $(A^2 + B^2)^{1/2}$ or, letting $\theta$ denote the angle between sides $A$ and $H$, set

$$A = H \cos\theta \tag{9.4}$$

and

$$B = H \sin\theta. \tag{9.5}$$

[3]The covariance matrix, of order $n_A + n_B + n_H$, can be built from

1. $\mathrm{var}(a_i) = \mathrm{var}(b_i) = \mathrm{var}(h_i) = u^2_{\rm rep} + u^2(\Delta L)$,
2. All covariances are equal to $u^2(\Delta L)$.

(The standard uncertainty $u(\Delta L)$ associated with the instrumental systematic effects may not be available explicitly from the calibration certificate of the instrument, but should be part of the detailed "uncertainty budget" for the calibration.)

The latter choice gives the least-squares formulation

$$\begin{aligned}\min_{H,\theta,\Delta L} S &= \sum_{i=1}^{n_A}\left(\frac{a_i - H\cos\theta - \Delta L}{u_{\text{rep}}}\right)^2 \\ &+ \sum_{i=1}^{n_B}\left(\frac{b_i - H\sin\theta - \Delta L}{u_{\text{rep}}}\right)^2 \\ &+ \sum_{i=1}^{n_H}\left(\frac{h_i - H - \Delta L}{u_{\text{rep}}}\right)^2.\end{aligned}$$

Its solution could be found using the Gauss-Newton algorithm or one of its variants [22].[4]

Many such problems would be solved in this manner. In this particular case, by defining new parameters

$$v_1 = H\cos\theta + \Delta L, \qquad v_2 = H\sin\theta + \Delta L, \qquad v_3 = H + \Delta L \tag{9.6}$$

and $\mathbf{v} = (v_1,\ v_2,\ v_3)^{\mathrm{T}}$, the problem becomes

$$\min_{\mathbf{v}} \sum_{i=1}^{n_A}(a_i - v_1)^2 + \sum_{i=1}^{n_B}(b_i - v_2)^2 + \sum_{i=1}^{n_H}(h_i - v_2)^2.$$

The problem separates into three trivial independent minimization problems, giving the solution

$$v_1 = \bar{a} = \sum_{i=1}^{n_A} a_i/n_A,$$

and similarly $v_2 = \bar{b}$ and $v_3 = \bar{h}$.

The equations (9.6) can then straightforwardly be solved for $H$, $\Delta L$ and $\theta$, from which $A$ and $B$ are determined from (9.4) and (9.5). The relevant uncertainties and covariance matrices are readily found using the principles of Section 6.

Since the value of $u_{\text{rep}}^2$ is common to all terms in the sum, the minimizing values of $H$, $\theta$ and $\Delta L$ do not depend on it. The term may therefore may replaced by unity (or any other constant). Thus, the solution can be obtained without knowledge of the random repeatability uncertainty or the systematic instrumental uncertainty.

The covariance matrix associated with the solution values of $H$, $\theta$ and $\Delta L$ provides the required output uncertainties.

## 9.5 Limit of detection

This example is intended to provide a simple illustration of how measurements of analyte concentration at the limit of detection can be analysed to furnish a value

[4] Advantage can be taken as follows of the fact that the problem is linear in two of the unknowns, $H$ and $\Delta L$. Equate to zero the partial derivatives of $S$, with respect to $H$, $\theta$ and $\Delta L$, to give three algebraic equations. Eliminate $H$ and $\Delta L$ to give a single nonlinear equation in $\theta$. Solve this equation using a suitable "zero-finder" (cf. Section 6.2.3). Determine $H$ and $\Delta L$ by substitution.

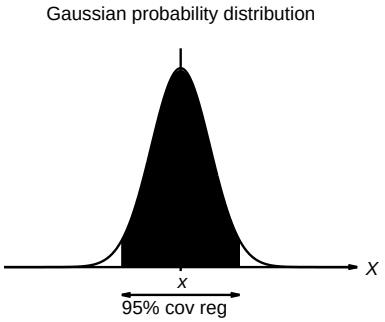


Figure 9.7: The Gaussian probability density function for the input quantity for the limit of detection problem. The 95% coverage interval that would conventionally be obtained for the analyte concentration is indicated.

for the measurand and its uncertainty. It utilizes basic statistical modelling principles.

The framework is as given in Section 4.2.2 on constrained uncertainty evaluation. The GUM model (4.5), repeated here, is

$$Y = \max(X,\ 0),$$

where the input quantity $X$ is observed (unconstrained) analyte concentration and the measurand $Y$ real (constrained) analyte concentration.

$X$ is estimated by $x$, the arithmetic mean of a number of (unconstrained) observations of analyte concentration. Its standard uncertainty is estimated by the standard deviation of this value. At or near the limit of detection, some of the observations would be expected to take negative values. If the measurements related to an analytical blank [32, Clause F2.3], used subsequently to correct other results, on average as many negative as positive values might be expected. If the analyte was actually present a preponderance of positive over negative values would be expected. Numerical values to represent this latter situation are chosen. The treatment is general, however, and can readily be repeated for other numerical values, even including a negative value for the mean of the observations.

Suppose that nothing is known about the observations other than that they can be regarded as independently and identically distributed. The use of the Principle of Maximum Entropy would indicate that the input $X$ can be regarded as a Gaussian variable with the above mean and standard deviation. For illustrative purposes, take the mean $x = 1.0$ ppm and the standard deviation $u(x) = 1.0$ ppm.

The pdf of the input quantity[5] and the model are thus fully defined.

Figure 9.7 illustrates this Gaussian pdf and indicates the 95% coverage interval that would conventionally be obtained for the analyte concentration.

For the above numerical values the corresponding pdf is given in Figure 9.8.

The area to the left of the origin under the Gaussian pdf with mean $x$ and standard deviation $u(x)$ is $\Phi((0 - x)/u(x))$ (see Section 4.3.1). The fraction of

[5]Other pdf's can be entertained. The subsequent treatment might not be as simple as that here, but can be addressed using Monte Carlo Simulation or other methods as appropriate.

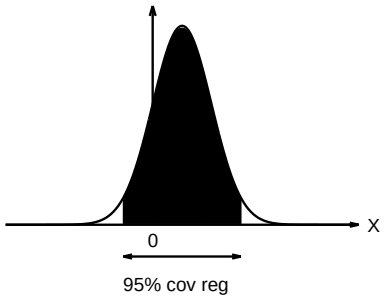


Figure 9.8: As Figure 9.7, but for the case in the text where the mean $x = 1.0$ ppm and the standard deviation $u(x) = 1.0$ ppm.

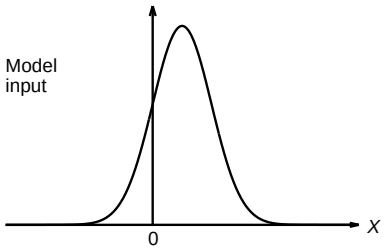


Figure 9.9: The pdf for the input quantity for the limit of detection problem

the values of $Y = \max(X,\ 0)$ that is zero is equal to this value. For the above numerical values, the fraction is $\Phi(-1.0) = 0.16$. So, 16% of the distribution of values that can plausibly be ascribed to the measurand take the value zero.

The pdf for the input quantity is indicated in Figure 9.9. That for the output quantity is depicted in Figure 9.10. In Figure 9.10 16% of the area under the curve is concentrated at the origin. Strictly, this feature should be denoted by a Dirac delta function (having "infinite height and zero width"). For illustrative purposes only, the function is depicted as a "tall thin" solid rectangle.

The shortest 95% coverage interval for the measurand therefore has (a) zero as its left-hand endpoint, and (b) as right-hand endpoint that value of $X$ for which $\Phi((X - x)/u(x)) = 0.95$, viz., $X = 2.6$ ppm. Thus, the required 95% coverage interval is $[0.0,\ 2.6]$ ppm.[6]

Figure 9.11 shows the distribution function $G(Y)$ for $Y$, with endpoints of the 95% coverage interval indicated by vertical lines. $G(Y)$ "rises instantaneously" at $Y = 0$ from zero to 0.16 and thereafter behaves as the Gaussian distribution function.[7]

The mean and standard uncertainty of $G(Y)$ are readily shown to be $y = 1.1$ ppm and $u(y) = 0.9$ ppm.

By comparison, the Mainstream GUM approach would yield a result as

[6] A Monte Carlo Simulation confirms this result.

[7] It is apparent from this figure that if a 95% coverage interval with 2.5% of the distribution in each tail were chosen the interval would be longer, in fact being [0.0, 3.0] ppm. Of course, since more than 5% of the distribution is at $Y = 0$, the left-hand endpoint remains at zero for *any* 95% coverage interval.

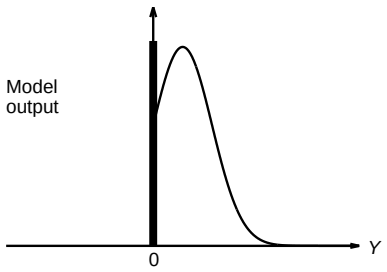


Figure 9.10: The pdf for the output quantity for the limit of detection problem. 16% of the area under the curve is concentrated at the origin. Strictly, this feature should be denoted by a Dirac delta function (having "infinite height and zero width"). For illustrative purposes the function is depicted as a "tall thin" solid rectangle.

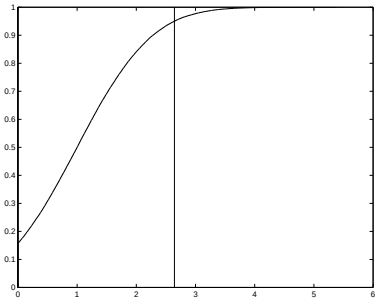


Figure 9.11: The distribution function for the limit of detection problem. The right-hand endpoint of the 95% coverage interval is indicated by a vertical line. the left-hand endpoint is at the origin.

follows. Since, in the neighbourhood of the input estimate $x = 1.0$ ppm, the model behaves as $Y = \max(X,\ 0) = X$, Mainstream GUM gives $y = 1.0$ ppm as a point estimate of the measurand. Moreover, the sensitivity coefficient $c$ is

$$\frac{\partial f}{\partial X} = 1,$$

evaluated at $X = 1.0$ ppm, viz., $c = 1$. Thus,

$$u(y) = |c|u(x) = 1.0 \text{ ppm}.$$

It follows that a 95% coverage interval based on Mainstream GUM considerations is $(1.0 \pm 2.0)$ ppm or $[-1.0,\ 3.0]$ ppm, following the use of Formula (6.2).

This interval is longer than the model-based interval $[0.0,\ 2.6]$ ppm and extends into the infeasible region. As stated earlier in this example, such an interval is appropriate for summarising the *observations*, but not for the *physically constrained measurand*, the real analyte concentration.

Similar principles can be applied to the measurement of the concentrations of a number of solution constituents. The analysis would be harder, but readily supported by the use of MCS.

## 9.6 Constrained straight line

The determination of suitable calibration lines and curves is a widespread requirement in metrology. The parameters of these lines and curves (and of models in general) may have to meet stipulated criteria in order that they reflect appropriate physical properties. For instance, a temperature in Kelvin cannot be negative.

Consider the length of a gauge block as its temperature is gradually increased. Suppose that for each of a sequence of increasing controlled temperature values the length of the block is measured. It is required to estimate the coefficient of expansion of the metal of which the block is made. The temperatures can be assumed to be known accurately and the measured lengths contain error. It is assumed that these errors have a Gaussian pdf.

A least-squares straight-line fit to the data is appropriate. The gradient of the line (the rate of change of length with respect to temperature) provides an estimate of the coefficient of expansion.

It can be shown that the gradient has an error that described by the Student's-$t$ distribution (cf. [1]). This distribution is the pdf of the output quantity and its use permits a coverage interval for the gradient to be obtained.

This process can often be expected to be satisfactory. This statement applies even though the Student's-$t$ distribution has infinite tails, implying that the left-hand tail includes zero and hence that there is a finite probability that the gradient is negative. This matter is of little concern since the tail probability is often very small indeed.

There are circumstances, however, where this aspect may be a concern, especially in the context of a procedure or computer software that might be used in a wide range of circumstances.

Consider a gauge block made from a material having a very small coefficient of expansion. In this situation the estimated uncertainty in the coefficient of

expansion could be comparable in size to the value of the coefficient itself. As a consequence, the application of conventional approaches to determining a coverage interval might produce an interval containing zero! Cf. the limit of detection example (Section 9.5).

Alternative approaches, including MCS, can be used to avoid this anomaly. Suppose that MCS is used to compute many estimates of the best-fitting straight line and hence of the gradient (expansion coefficient). Each MCS trial involves sampling from the Gaussian pdf associated with the measured lengths, fitting a *constrained* line to the sampled data given by these lengths corresponding to the fixed values of the independent variables, and taking its gradient as the corresponding measurement result. The set of gradients so obtained form the basis, as in Section 9.5, for a distribution function of the gradient and hence a coverage interval.

The term "constrained line" is used to indicate the fact that for any set of sampled data a straight line with an intercept parameter and a gradient parameter must be fitted subject to the condition that the gradient is not negative. It is straightforward to use conventional fitting procedures for this purpose. First, a straight line is fitted to the data without imposing the condition. If the gradient of the line were positive (or zero) the line would automatically satisfy the condition and would therefore be the required solution. Otherwise, the "best" line that can be fitted that satisfies the constraint would have a zero gradient. Such a line is a constant. This constant is easily found, since the best least-squares fit by a constant is the same problem as finding the arithmetic mean of the data.

Thus the sequence of $M$, say, gradients so obtained will include some zero values, the remainder being strictly positive. The estimated distribution function for the gradient $Y$, as that for the limit of detection problem (Section 9.5), therefore has a jump discontinuity at $y = 0$, the magnitude of which is the proportion of trials that gave zero gradient, followed by a "smooth" increase through increasing gradient values.

A simple Monte Carlo Simulation was carried out. The data used consisted of the points (18, 23), (19, 24), (20, 26), (21, 27), (22, 28), where the first coordinated denotes temperature in °C and the second length measurement in a normalised variable. Figure 9.12 depicts the five gauge block length measurements against temperature (large blobs) and 100 simulations (small blobs) of these measurements obtained from sampling from assigned Gaussian pdf's. For the simulation the standard uncertainty of the length measurements was taken as 0.005 and used as the standard deviation of the Gaussian pdf's.

For some of the synthesised sets of five measurements the gradient of an unconstrained least-squares straight line would be negative, were it not infeasible.

The results from the use of a large number (100,000) of trials gave a distribution function for the gradient very similar to that for the limit of detection problem (Section 9.5).

## 9.7 Fourier transform

Consider measurements of a periodic phenomenon. Such measurements are commonplace in many branches of metrology. Suppose a complete period is

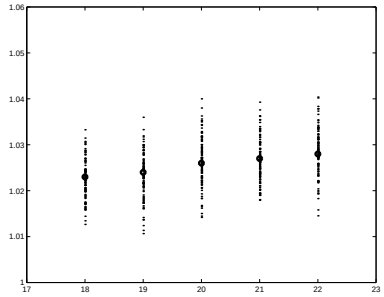


Figure 9.12: The five gauge block length measurements against temperature (large blobs) and 100 simulations of these measurements obtained from sampling from assigned Gaussian pdf's. For some of the synthesised sets of five measurements the gradient of an unconstrained least-squares straight line would be negative.

measured in that $n$ values $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ are available. These values correspond to the uniformly spaced angles $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$, where $\theta_i = 2\pi(i-1)/n$.

A Fourier transform of such data provides information concerning the frequency content of the data.

Each Fourier coefficient depends on all (or most of) the values of $x_i$, regarded as the input quantities, and is a linear combination of them.

Suppose that the errors of measurement can be regarded as mutually independent with standard deviation $\sigma$. The covariance matrix for the measurement errors or, equivalently, for the input quantities $x_i$, is therefore given by

$$V_{\mathbf{x}} = \sigma^2 I, \tag{9.7}$$

where $I$ is the identity matrix of order $n$. It is required to evaluate the uncertainties in the Fourier transform of this data, i.e., in the coefficients of the Fourier representation of the data. The coefficients constitute the (vector) measurand.

The Fourier *representation* of the data is

$$h(\theta) = a_0 + a_1 \cos\theta + b_1 \sin\theta + \cdots + a_r \cos r\theta + b_r \sin r\theta,$$

where $r = \lfloor n/2 \rfloor$. (When $n$ is even, the coefficient $b_r$ of $\sin r\theta$ is in fact zero.) Let $\mathbf{y} = (y_1, \ldots, y_{2r+1})^{\mathrm{T}} = (a_0, a_1, b_1, \ldots, a_r, b_r)^{\mathrm{T}}$, the output quantities. The Fourier transform $\mathbf{y}$ of $\mathbf{x}$ is then given implicitly by

$$\mathbf{x} = A\mathbf{y}, \tag{9.8}$$

where

$$A = \begin{bmatrix} 1 & \cos\theta_1 & \sin\theta_1 & \ldots & \cos r\theta_1 & \sin r\theta_1 \\ 1 & \cos\theta_2 & \sin\theta_2 & \ldots & \cos r\theta_2 & \sin r\theta_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cos\theta_n & \sin\theta_n & \ldots & \cos r\theta_n & \sin r\theta_n \end{bmatrix}$$

is the matrix of order $n$ of Fourier basis-function values. Formally, the Fourier coefficients are given in terms of the data using

$$\mathbf{y} = A^{-1}\mathbf{x} \tag{9.9}$$

or, equivalently, from a formula that expresses the $t_i$ as linear combinations of the $x_i$, where the multipliers are sine and cosine terms. (In practice, $\mathbf{t}$ would be computed from $\mathbf{x}$ using the fast Fourier transform (FFT) [8].) The FFT gives far greater efficiency than would be obtained from the application of general-purpose linear-algebra techniques, and also greater numerical accuracy.[8]

Denote the covariance matrix of $\mathbf{y}$ by $V_{\mathbf{y}}$. The application of (6.10) to (9.8) gives

$$V_{\mathbf{x}} = AV_{\mathbf{y}}A^{\mathrm{T}}.$$

This result is *exact* since the output quantity $\mathbf{y}$ and the input quantity $\mathbf{x}$ are related linearly through (9.8), and linearisation introduces no error in this case. Since $A$ is invertible,

$$V_{\mathbf{y}} = A^{-1}V_{\mathbf{x}}A^{-\mathrm{T}}.$$

This expression enables in general the covariance of the Fourier coefficients to be computed from that of the data. As a consequence of (9.7),

$$V_{\mathbf{y}} = \sigma^2 A^{-1}A^{-\mathrm{T}} = \sigma^2(A^{\mathrm{T}}A)^{-1}.$$

Now, using the fact that $\boldsymbol{\theta}$ is equiangular and the fundamental properties of the trigonometric functions, it is straightforward to show that

$$A^{\mathrm{T}}A = \frac{n}{2}\mathrm{diag}\left\{\ 2,\quad 1,\quad \ldots,\quad 1\ \right\},$$

giving

$$(A^{\mathrm{T}}A)^{-1} = \frac{2}{n}\mathrm{diag}\left\{\ \tfrac{1}{2},\quad 1,\quad \ldots,\quad 1\ \right\}.$$

Consequently,

$$V_{\mathbf{y}} = \frac{2}{n}\sigma^2\mathrm{diag}\left\{\ \tfrac{1}{2},\quad 1,\quad \ldots,\quad 1\ \right\}.$$

This result states that for data errors that are independent with standard uncertainty $\sigma$, the errors in the Fourier coefficients are (also) independent, with standard uncertainty equal to $\sigma$ scaled by the factor $\sqrt{2/n}$, where $n$ is the number of points (with the exception of the constant term for which the factor is $\sqrt{1/n}$). Moreover, each Fourier coefficient is a linear combination of $n$ raw data values, the multipliers being the products of a constant value and that of values of cosines and sines (and thus lying between -1 and +1). Consequently, if $n$ is large[9] regardless of the statistical distributions of the errors in the data, the errors in the Fourier coefficients can be expected to be very close to being normally distributed. This result is an immediate consequence of the Central Limit Theorem when using the Fourier transform to analyse large numbers of measurements having independent errors. Thus, it is valid to regard the resulting Fourier coefficients as if they were *independent Gaussian-distributed measurements*.

The outputs, the Fourier coefficients, from this process become the inputs to the subsequent stage, viz., the evaluation of the Fourier series $h(\theta)$ for any value of $\theta$. Now, since, as shown, the Fourier coefficients are uncorrelated,

$$u^2(h(\theta)) = u^2(a_0)+u^2(a_1)\cos^2\theta+u^2(b_1)\sin^2\theta\cdots+u^2(a_r)\cos^2 r\theta+u^2(b_r)\sin^2 r\theta.$$

[8] In exact arithmetic, the FFT and (9.9) give identical results, since mathematically they are both legitimate ways of expressing the solution.

[9] In high-accuracy roundness measurement, e.g., $n = 2000$ would be typical.

Using the results above,

$$u^2(h(\theta)) = \frac{\sigma^2}{n} + \frac{2\sigma^2}{n}\cos^2\theta + \frac{2\sigma^2}{n}\sin^2\theta + \ldots + \frac{2\sigma^2}{n}\cos^2 r\theta + \frac{2\sigma^2}{n}\sin^2 r\theta, \quad (9.10)$$

which simplifies to $\sigma^2$. Thus,

$$u(h(\theta)) = \sigma,$$

i.e., the uncertainty in the Fourier representation of a data set when evaluated at any point is identical to the uncertainty in the data itself. This property is remarkable in that the (interpolatory) replacement of data by other functions usually gives an amplification of the raw data uncertainty, at least in some regions of the data.

# Chapter 10

# Recommendations

The Guide to the Expression of Uncertainty in Measurement (GUM) provides internationally-agreed recommendations for the evaluation of uncertainties in metrology. Central to the GUM is a measurement model with input quantities, defined by probability distributions, and a measurand with a corresponding measurement result that, consequently, is also a probability distribution. The use of the GUM permits the uncertainty of the measurand to be evaluated. In particular, an interval (termed here a coverage interval) that can be expected to encompass a specified fraction of the distribution of values that could reasonably be attributed to the measurand can be obtained.

It will always be necessary to make some assertions about the uncertainties associated with the model input quantities. That is the metrologist's task. To state no knowledge of certain contributions is unhelpful. The metrologist needs to make statements, using expert judgement if necessary, about what he believes, and those statements provide the basis for the analysis, until better statements become available. After all, he is best-placed to do this. If everything is recorded, the quoted uncertainty can be defended in that light.

Arguably, the worst-case scenario is when the metrologist genuinely feels he knows nothing about the nature of an uncertainty contribution other than an asserted upper limit on the uncertainty of the input quantity. (If he cannot even quote that, the uncertainty evaluation cannot be progressed at all!) In this situation the Principle of Maximum Entropy would imply that the best estimate of the underlying probability distribution is uniform, with bounds provided by the limit.

In general, it is recommended that all model input quantities are characterised in terms of pdf's. By doing so the metrologist is able to incorporate to the maximum his degree of belief in the various input quantities. In particular, if little or very little information is available, appeal to the Principle of Maximum Entropy permits a defensible pdf to be provided.

Once the model, and the pdf's of the input quantities are in place, it is possible to use a variety of tools for determining the pdf of the output quantity and thence a coverage interval or coverage region.

The attributes of the various approaches considered, all in a sense covered by the full GUM document, are to be taken into account when selecting whether to apply Mainstream GUM, MCS or other analytical or numerical methods.

Validation of the approach used is important in cases of doubt. The use of

MCS to validate Mainstream GUM is urged when it is unclear whether the latter is applicable in a certain situation. MCS can also be seen as a widely-applicable tool for uncertainty evaluation.

> [GUM, Clause 0.4] The actual quantity used to express uncertainty should be:
>
> 1. Internally consistent: it should be directly derivable from the components that contribute to it, as well as independent of how these components are grouped and of the decomposition of the components into subcomponents;
> 2. Transferable: it should be possible to use directly the uncertainty evaluated for one result as a component in evaluating the uncertainty of another measurement in which the first result is used.
> 3. ...it is often necessary to provide an interval about the measurement result that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the quantity subject to measurement. Thus the ideal method for evaluating and expressing uncertainty in measurement should be capable of readily providing such an interval, in particular, one with a coverage probability or level of probability that corresponds in a realistic way with that required.

These are laudable properties and objectives. It is reasonable to summarize them and to infer further aims as follows:

1. All information used to evaluate uncertainties is to be recorded.
2. The sources of the information are to be recorded.
3. Any assumptions or assertions made are to be stated.
4. The model and its input quantities are to be provided in a manner that maximizes the use of this information consistent with the assumptions made.
5. Uncertainties are to be evaluated in a manner that is consistent with quality management systems and, in particular, the results of the evaluation are to be fit for purpose.
6. If the same information is provided to different bodies, the uncertainties these bodies calculate for the required results are to agree to within a stipulated numerical accuracy.
7. Difficulties in handling sparse or scarce information are to be addressed by making alternative, equally plausible assumptions, and re-evaluating the uncertainty of the measurand to obtain knowledge of the variability due to this source.

The intention of this guide has been to address these aims as far as reasonably possible. Further, the two-phase approach advocated in Section 3.2 and followed in the rest of this guide supports the first point from GUM, Clause 0.4. The approach to multi-stage models recommended here supports the second point. Finally, the mathematical formulation and the attitude of this guide supports the third point, through the solution approaches of Section 5.

# Bibliography

[1] Guide to the Expression of Uncertainty in Measurement. International Organisation for Standardisation, Geneva, Switzerland, 1995. ISBN 92-67-10188-9, Second Edition. 1, 3, 9, 13, 3.2, 4.2, 20, 24, 6.1, 65, 9.6, 81, 84

[2] AFNOR. Aid in the procedure for estimating and using uncertainly in measurements and test results. Technical Report FD X 07-021, AFNOR, 2000. Informal English translation. 4.2.1

[3] V. Barnett. The ordering of multivariate data. *J. R. Statist. Soc. A*, 139:318, 1976. 62

[4] M. Basil and A. W. Jamieson. Uncertainty of complex systems by Monte Carlo Simulation. In *North Sea Flow Measurement Workshop, Gleneagles, 26-29 October 1998*, 1998. 5.3.1

[5] Stephanie Bell. Measurement Good Practice Guide No. 11. A Beginner's Guide to Uncertainty of Measurement. Technical report, National Physical Laboratory, 1999. 4.2

[6] P. Mac Berthouex and L. C. Brown. *Statistics for Environmental Engineers.* CRC Press, USA, 1994. 65

[7] BIPM. Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes. Technical report, Bureau International des Poids et Mesures, Sèvres, France, 1999. 2.2

[8] E. O. Brigham. *The Fast Fourier Transform.* Prentice-Hall, New York, 1974. 9.7

[9] B. P. Butler, M. G. Cox, A. B. Forbes, P. M. Harris, and G. J. Lord. Model validation in the context of metrology: A survey. Technical Report CISE 19/99, National Physical Laboratory, Teddington, UK, 1999. 2, 9

[10] P. J. Campion, J. E. Burns, and A. Williams. A code of practice for the detailed statement of accuracy. Technical report, National Physical Laboratory, Teddington, UK, 1980. 26, 7.7.1

[11] G. Casella and R. L. Berger. *Statistical Inference.* Wadsworth and Brooks, Pacific Grove, Ca., USA, 1990. 2.2

[12] P. Ciarlini, M. G. Cox, E. Filipe, F. Pavese, and D. Richter, editors. *Advanced Mathematical and Computational Tools in Metrology V. Series on Advances in Mathematics for Applied Sciences Vol. 57*, Singapore, 2001. World Scientific. 1.4

[13] P. Ciarlini, M. G. Cox, R. Monaco, and F. Pavese, editors. *Advanced Mathematical Tools in Metrology*, Singapore, 1994. World Scientific. 1.4

[14] P. Ciarlini, M. G. Cox, F. Pavese, and D. Richter, editors. *Advanced Mathematical Tools in Metrology II*, Singapore, 1996. World Scientific. 1.4

[15] P. Ciarlini, M. G. Cox, F. Pavese, and D. Richter, editors. *Advanced Mathematical Tools in Metrology III*, Singapore, 1997. World Scientific. 1.4

[16] P. Ciarlini, A. B. Forbes, F. Pavese, and D. Richter, editors. *Advanced Mathematical and Computational Tools in Metrology IV*, Singapore, 2000. World Scientific. 1.4

[17] Analytical Methods Committee. Uncertainty of measurement: implications of its use in analytical science. *Analyst*, 120:2303–2308, 1995. 13

[18] M. G. Cox. The numerical evaluation of B-splines. *Journal of the Institute of Mathematics and its Applications*, 8:36–52, 1972. 1

[19] M. G. Cox, M. P. Dainton, A. B. Forbes, P. M. Harris, P. M. Schwenke, B. R. L Siebert, and W. Wöger. Use of Monte Carlo Simulation for uncertainty evaluation in metrology. In P. Ciarlini, M. G. Cox, E. Filipe, F. Pavese, and D. Richter, editors, *Advanced Mathematical Tools in Metrology V. Series on Advances in Mathematics for Applied Sciences Vol. 57*, pages 93–104, Singapore, 2001. World Scientific. 17, 5.3.1, 57

[20] M. G. Cox, M. P. Dainton, and P. M. Harris. Software specifications for uncertainty evaluation and associated statistical analysis. Technical Report CMSC 10/01, National Physical Laboratory, Teddington, UK, 2001. 5, 52, 53, 7.1.2, 7.7, 7.8, 8.2

[21] M. G. Cox, M. P. Dainton, P. M. Harris, and N. M. Ridler. The evaluation of uncertainties in the analysis of calibration data. In *Proceedings of the 16th IEEE Instrumentation and Measurement Technology Conference*, pages 1093–1098, Venice, 24-26 May 1999, 1999. 5.3.1

[22] M. G. Cox, A. B. Forbes, and P. M. Harris. Software Support for Metrology Best Practice Guide No. 4. discrete modelling. Technical report, National Physical Laboratory, Teddington, UK, 2000. 9, 4.1, 4.1, 20, 4, 24, 7, 74, 74

[23] M. G. Cox and P. M. Harris. Up a GUM tree? *The Standard*, 99:36–38, 1999. 5.3.1

[24] M. G. Cox and E. Pardo. The total median and its uncertainty. In P. Ciarlini, M. G. Cox, E. Filipe, F. Pavese, and D. Richter, editors, *Advanced Mathematical Tools in Metrology V. Series on Advances in Mathematics for Applied Sciences Vol. 57*, pages 105–116, Singapore, 2001. World Scientific. 54

[25] T. J. Dekker. Finding a zero by means of successive linear interpolation. In B. Dejon and P. Henrici, editors, *Constructive Aspects of the Fundamental Theorem of Algebra*, London, 1969. Wiley Interscience. 6.2.3

[26] C. F. Dietrich. *Uncertainty, Calibration and Probability.* Adam Hilger, Bristol, UK, 1991. 28, 5.1, 5.1.1

[27] DIN. DIN 1319 Teil 4 1985 Grundbegriffe der Meßtechnik; Behandlung von Unsicherheiten bei der Auswertung von Messungen, 1985. 10

[28] EA. Expression of the uncertainty of measurement in calibration. Technical Report EA-4/02, European Co-operation for Accreditation, 1999. 2.2, 9, 5.1.1, 5.1.2, 5.1.2, 67, 67, 9.3, 9.3

[29] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1982. 53

[30] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* Monographs on Statistics and Applied Probability 57. Chapman and Hall, 1993. 53, 57

[31] G. F. Engen. *Microwave Circuit Theory and Foundations of Microwave Metrology.* Peter Peregrinus, London, 1992. 50

[32] Eurachem. Quantifying uncertainty in analytical measurement. Technical report, Eurachem, 2000. Final draft. 4.2.2, 9.5

[33] M. J. Fryer. A review of some non-parametric methods of density estimation. *J. Inst. Math. Applic.*, 20:335–354, 1977. 8

[34] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization.* Academic Press, London, 1981. 6.2.3, 6.2.4

[35] G. H. Golub and C. F. Van Loan. *Matrix Computations.* John Hopkins University Press, Baltimore, MD, USA, 1996. Third edition. 4, 8

[36] B. D. Hall. Computer modelling of uncertainty calculations with finite degrees of freedom. *Meas. Sci. Technol.*, 11:1335–1341, 2000. 14, E

[37] B. D. Hall and R. Willink. Does Welch-Satterthwaite make a good uncertainty estimate? *Metrologia*, 2000. In press. 3

[38] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods.* Wiley, New York, 1964. 37

[39] ISO. ISO Guide 35 - Certification of reference materials - general and statistical principles. International Organisation for Standardisation, Geneva, Switzerland, 1989. 10

[40] ISO. International Vocabulary of Basic and General Terms in Metrology. International Organisation for Standardisation, Geneva, Switzerland, 1993. B

[41] ISO. ISO 3534-1:1993. Statistics - Vocabulary and symbols - Part 1: Probability and general statistical terms. International Organisation for Standardisation, Geneva, Switzerland, 1993. A.3

[42] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev*, 106:620–630, 1957. D.2, D.2

[43] E. T. Jaynes. Where do we stand on maximum entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, Cambridge, Ma., 1978. MIT Press. D.2

[44] R. Kacker. An interpretation of the Guide to the Expression of Uncertainty in Measurement. In *NCSL-2000 proceedings, Toronto*, 2000. D.1

[45] D. M. Kerns and R. W. Beatty. *Basic Theory of Waveguide Junctions and Introductory Microwave Network Analysis*. Pergamon Press, London, 1967. 50

[46] T. Leonard and J. S. J. Hsu. *Bayesian Methods*. Cambridge University Press, Cambridge, 1999. 5

[47] S. Lewis and G. Peggs. *The Pressure Balance: A Practical Guide to its Use. Second edition*. HMSO, London, 1991. 6, 47

[48] National Physical Laboratory, *http://www.npl.co.uk/ssfm/index.html. Software Support for Metrology Programme*. 1.1

[49] S. D. Phillips, W. T. Estler, M. S. Levenson, and K. R. Evehardt. Calculation of measurement uncertainty using prior information. *J. Res. Natl. Inst. Stnds. Technol.*, 103:625–632, 1998. 3, 85

[50] D. Rayner. Overview of the UK Software Support for Metrology Programme. In P. Ciarlini, A. B. Forbes, F. Pavese, and D. Richter, editors, *Advanced Mathematical and Computational Tools in Metrology IV*, pages 197–201, Singapore, 2000. World Scientific. 1.1

[51] J. R. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, Ca., USA, second edition, 1995. 4.3, 5.1.1, 37, 5, 62, A.1

[52] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 1999. 12

[53] J. G. Saw, M. C. K. Yang, and T. C. Mo. Chebychev's inequality with estimated mean and variance. *American Statistician*, 38:130–132, 1984. 4.6.1

[54] H. Schwenke, B. R. L. Siebert, F. Wäldele, and H. Kunzmann. Assessment of uncertainties in dimensional metrology by Monte Carlo simulation: proposal of a modular and visual software. *Annals CIRP*, 49:395–398, 2000. 5.3.1

[55] R. S. Scowen. QUICKERSORT, Algorithm 271. *Comm. ACM*, 8:669, 1965. 11

[56] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois, Urbana-Champagne, 1949. D.2

[57] B. R. L. Siebert. Discussion of methods for the assessment of uncertainties in Monte Carlo particle transport calculations. In *Advanced Mathematical and Computational Tools in Metrology IV*, pages 220–229, 2000. 37

[58] C. G. Small. A survey of multidimensional medians. *International Statistical Review*, 58:263, 1990. 62

[59] C. R. Smith and W. T. Grandy, editors. *Maximum-Entropy and Bayesian methods in inverse problems*, Dordrecht, 1985. Reidel. D.2

[60] P. I. Somlo and J. D. Hunter. *Microwave Impedance Measurement*. Peter Peregrinus, London, 1985. 50

[61] B. N. Taylor and C. E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical Report TN1297, National Institute of Standards and Technology, 1994. 9, 10, 3.2, 7.8

[62] UKAS. The treatment of uncertainty in EMC measurements. Technical Report NIS 81, United Kingdom Accreditation Service, 1994. 5.1.1, B

[63] UKAS. The expression of uncertainty and confidence in measurement. Technical Report M 3003, United Kingdom Accreditation Service, 1997. 2.2

[64] P. A. G. van der Geest. An algorithm to generate samples of multi-variate distributions with correlated marginals. *Comput. Stat. Data Anal.*, 27:271–289, 1998. 5.3.1, 52, 53

[65] K. Weise and W. Wöger. A Bayesian theory of measurement uncertainty. *Measurement Sci. Technol.*, 3:1–11, 1992. 13, 4.3.5, 81, D.2

[66] W. Wöger. Probability assignment to systematic deviations by the Principle of Maximum Entropy. *IEE Trans. Instr. Measurement*, IM-36:655–658, 1987. 81, D.2, D.2

# Appendix A

# Some statistical concepts

Some statistical concepts used in this guide are reviewed. The concept of a random variable is especially important. Inputs and outputs are (or are to be regarded as) random variables. Some of the elementary theory of random variables is pertinent to the subsequent considerations.

## A.1 Discrete random variables

A *discrete random variable* $X$ is a variable that can take only a finite number of possible values. If $X$ is the number of heads in an experiment consisting of tossing three coins, $X$ can take (only) the value 0, 1, 2 or 3.

A *frequency function* states the probabilities of occurrence of the possible outcomes. For the coin-tossing experiment, the probability $P(X)$ that the outcome is $X$ is given by

$$\begin{aligned} P(X=0) &= \frac{1}{8}, \\ P(X=1) &= \frac{3}{8}, \\ P(X=2) &= \frac{3}{8}, \\ P(X=3) &= \frac{1}{8}. \end{aligned}$$

The probabilities can be deduced by enumerating all $2 \times 2 \times 2 = 8$ possible outcomes arising from the fact that each coin can only land in one of two equally likely ways, or by using the binomial distribution [51, p36] that applies to the analysis of such probability problems.

The *distribution function* $G(x)$ gives the probability that a random variable takes a value no greater than a specified value:

$$G(x) = P(X \leq x), \qquad -\infty < x < \infty.$$

For the coin-tossing experiment,

$$\begin{aligned} G(x<0) &= 0, \\ G(0 \leq x < 1) &= \frac{1}{8}, \end{aligned}$$

$$\begin{aligned} G(1 \leq x < 2) &= \frac{1}{2}, \\ G(2 \leq x < 3) &= \frac{7}{8}, \\ G(3 \leq x) &= 1. \end{aligned}$$

The distribution function varies from zero to one throughout its range, never decreasing.

The probability that $X$ lies in an interval $[a,\ b]$ is

$$P(a \leq X \leq b) = G(b) - G(a).$$

Two important statistics associated with a discrete random variable are its mean and standard deviation.

Let $x_1, x_2, \ldots$ denote the possible values of $X$ and $p(x)$ the frequency function of $X$. The expected value or *mean* $\mu$ of a discrete random variable $X$ with frequency function $p(x)$ is

$$\mu = \sum_i x_i p(x_i).$$

It is a measure of the location of $X$.

The *standard deviation* $\sigma$ of a discrete random variable $X$ with frequency function $p(x)$ is the square root of the *variance*.

$$\text{var}(X) = \sigma^2 = \sum_i (x_i - \mu)^2 p(x_i).$$

It is a measure of the spread or dispersion of $X$.

## A.2 Continuous random variables

A *continuous random variable* $X$ is a variable that can take any value within a certain interval. For a machine capable of weighing any person up to 150 kg, the recorded weight $x$ can take any value in the interval 0 kg to 150 kg.

For a continuous random variable $X$, the counterpart of the frequency function (for a discrete random variable) is the *probability density function* (pdf) $g(x)$. This function has the property that the probability that $X$ lies between $a$ and $b$ is

$$P(a < X < b) = \int_a^b g(x)dx.$$

Since a random variable $X$ must take *some* value, $g(x)$ has the property that

$$\int_\infty^\infty g(x)dx = 1.$$

The uniform density function is a density function that describes the fact that $X$ is equally likely to lie anywhere in an interval $[a,\ b]$:

$$g(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

The *distribution function* $G(x)$ gives the probability that a random variable takes a value no greater than a specified value, and is defined as for a discrete random variable:

$$G(x) = P(X \leq x), \qquad -\infty < x < \infty.$$

The distribution function can be expressed in terms of the probability density function as

$$G(x) = \int_{-\infty}^{x} g(v)dv.$$

The mean of a continuous random variable $X$ with probability density function $g(x)$ is

$$E(X) = \int_{-\infty}^{\infty} xg(x)dx.$$

It is often denoted by $\mu$.

The variance of a continuous random variable $X$ with density function $g(x)$ and expected value $\mu = E(X)$ is

$$\text{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 g(x)dx.$$

The variance is often denoted by $\sigma^2$ and its square root is the standard deviation $\sigma$.

## A.3 Coverage interval

A *coverage interval* (or *statistical coverage interval*) is an interval for which it can be stated with a given level of probability that it contains at least a specified proportion of the population [41].

Given the pdf $g(x)$, with distribution function $G(x)$, of a random variable $X$, the $(100p)$th percentile is the value $x_p$ such that

$$G(x_p) = \int_{-\infty}^{x_p} g(v)dv = p,$$

i.e., $(100p)\%$ of the pdf lies to the left of $x_p$.

A 95% coverage interval is therefore $[x_{0.025},\ x_{0.975}]$.

The inverse distribution $G^{-1}(p)$ permits the value of $X$ corresponding to a specified percentile to be obtained:

$$x_p = G^{-1}(p).$$

**Example 24** *A coverage interval for a Gaussian probability density function*

A 95% coverage interval for a Gaussian pdf with zero mean and unit standard deviation is $[-2.0, 2.0]$.

**Example 25** *A coverage interval for a uniform probability density function*

A 95% coverage interval for a uniform pdf with zero mean and unit standard deviation is $[-1.6, 1.6]$.

## A.4 95% versus $k = 2$

The use of a coverage probability of 95% has been adopted by many nations in quoting measurement results. There are two sides to the coin. In an interlaboratory comparison of measurement standards, an individual laboratory may lie more than two standard deviations from the mean (corresponding in a Gaussian world to an endpoint of a 95% coverage interval). There is a stigma attached to such a situation even though by chance alone such an "event" is likely to happen once in twenty.

Thus, there will be sensitivities at international level in dealing with comparisons in this regard (and also of course at all levels of the metrological chain). Consideration could therefore be given to the use of a coverage interval corresponding to a greater level of probability, e.g., 99% or 99.9%. Fewer laboratories or results would consequently be perceived as exceptional.

A similar interpretation arises within a calibration laboratory. Suppose that such a laboratory repeatedly makes calibrations for its customers, quoting realistic coverage intervals for the calibration results. On average one out of twenty of the provided calibration results will be outside the quoted coverage interval for the value of the measurand, on the basis of chance alone.

A major objection to providing a coverage interval corresponding to a high level of probability is that a sufficiently detailed knowledge of the tails of the distribution of values that can be attributed to the measurand is required. Such information is much harder to obtain, necessitating an appreciably larger number of measurements, for example, to quantify the distributions of the model inputs. Obtaining such information might be justified in safety-critical applications, where "risk of failure" has severe consequences.

The use of the 95% level of probability can be regarded as a matter of convenience. It is not perfect, but criticism could be directed at any level for reasons that are either general or specific to that level. A very large number of practitioners employ it, and the inertia that has been generated would imply that it should only be changed if there were exceptional reasons for doing so.

There is another important consideration relating to the choice of a 95% coverage interval.

A culture exists in certain metrological quarters in which, when determining or interpreting coverage intervals, "95%" is regarded as identical to "$k = 2$".[1] However, the use of $k = 2$ *in isolation* imparts no knowledge or assumption relating to the underlying statistical distribution. It might be Gaussian. It might be something completely different. As a consequence, different statements, each containing an uncertainty quoted in this way, cannot readily be compared because they might be associated with different levels of probability. In practice, quoting $k = 2$ (alone) would often be interpreted as relating to a Gaussian distribution and hence be equivalent to the 95% level of probability. This is a big assumption. The converse of this statement is that there is a great temptation simply to evaluate the standard uncertainty (standard deviation) of a measurement result (in some way) and double it in order to obtain a $k = 2$ value, which indeed it is, but at what price? The evaluation of uncertainties, particularly the Type B contributions (obtained by non-statistical techniques [1]), is far from

[1] The value of $k$, in the language of the GUM, is the coverage factor. It is used to scale the standard deviation of the pdf of the output, to obtain a coverage interval at the stated level of probability.

an easy task. However, whenever an uncertainty evaluation is carried out, the practitioners involved should make and record their knowledge and assertions concerning the statistical distributions (hypothesized if necessary) associated with all the inputs. Then, armed with a measurement model, which the GUM mandates, the propagation of these distributions through the model would imply a distribution for the measurement result, from which a 95% interval can be obtained. Mainstream GUM assumes the validity of the Central Limit Theorem and the linearizability of the model, the implication of which is that the measurement result follows a Gaussian or a $t$-distribution, the latter applying if the number of degrees of freedom associated with the measurement result is finite. Even if Mainstream GUM is followed and the limitations inherent in its use for the current purposes are defensible, the coverage factor will differ from two if the number of degrees of freedom is not large. If, as is permitted by the GUM (Clause G.1.5), other analytical or numerical methods are used, a 95% coverage interval can be obtained, as implied above (using Monte Carlo Simulation, for example), free of these limitations. When these other approaches are employed, there is no concept of a coverage factor. It will always be necessary to make some assertions about the uncertainties associated with the influence factors, but that is the metrologists task. To throw ones hands in the air and state no knowledge of certain contributions is unhelpful. The metrologist needs to make statements about what he believes, and those statements provide the basis for the analysis, until better statements become available. After all, he is best-placed to do this. If everything is recorded, the quoted uncertainty can be defended in that light. Arguably, the worst-case scenario is when the metrologist knows nothing about the nature of an uncertainty contribution than an asserted upper limit on the uncertainty of the corresponding input. In this situation the Principle of Maximum Entropy, as advocated by the Bayesians [66, 65], would imply that the best estimate of the underlying statistical distribution is uniform, with bounds specified by the limit. The Frequentists would have nothing to offer because no measured data is available! See Section D.

# Appendix B

# A discussion on modelling

In order to provide an objective evaluation of uncertainty, a model, despite the possibly great difficulties in obtaining it, is required.

> [GUM Clause 8, Step 1] Express mathematically the relationship between the measurand $Y$ and the input quantities $X_i$ on which $Y$ depends: $Y = f(X_1, X_2, \ldots, X_N)$. The function $f$ should contain every quantity, including all corrections and correction factors, that can contribute a significant component of uncertainty to the result of measurement.

The GUM thus mandates a model, which in the simplest case might be a straightforward mathematical formula, and in a hard case a complicated procedure involving software calculations, experimental rigs, etc.

In terms of inputs $\mathbf{X}$, outputs (results of measurement) $\mathbf{Y}$ and a model $\mathbf{f}$, then, consistent with the GUM, $\mathbf{Y} = \mathbf{f}(\mathbf{X})$. Also, evidently, once the model is in place, the uncertainties in $\mathbf{X}$ can be propagated through the model to become uncertainties in $\mathbf{Y}$. Alternatively and more generally, the pdf's of $\mathbf{X}$ can be propagated through the model to yield the (possibly multidimensional) *pdf* of $\mathbf{Y}$.

These considerations are at the heart of uncertainty evaluation.

If no model is available, even if its inputs and the pdf's for these inputs have been provided, it is not possible to work with uncertainties in this way or, arguably, in any way. The United Kingdom Accreditation Service (UKAS) guidance [62] on obtaining uncertainties in the area of electromagnetic compatibility (EMC) is not based on the concept of a model. As a consequence, it is difficult to determine the extent to which the guidance is adequate, and to extend the presented considerations and examples to cover other circumstances, especially to account for further inputs, is not feasible. Of course, in *any* treatment of uncertainties, a model is implicitly assumed, even if not formally provided.

Additionally, as arise in a number of areas of metrology, the EMC guide [62] considers the effects of nonlinear changes of variable, specifically the need to work with linear variables and logarithmic variables, as a consequence of measurements expressed in decibel units. Such transformations arise in a number of areas of metrology. The handling of the uncertainties that "propagate" through these transformations is handled in a heuristic manner. The approach is adequate when the uncertainties are sufficiently small. However, in the EMC

area and in some other branches of metrology, this is not always the case. Not only can such a heuristic approach give rise to invalid results for the propagated uncertainty, it can also yields anomalies resulting from the fact that the propagation of uncertainties associated with a symmetric pdf such as a Gaussian or a uniform pdf through a nonlinear transformation yields an asymmetric distribution. The extent of the asymmetry is exaggerated by large input uncertainties.

UKAS has recognised the problems associated with the current guidance document and is working towards a revision that will incorporate a number of the considerations included in this best-practice guide. These considerations include a model-based approach, handling changes of variables in a valid manner, the use of Monte Carlo Simulation to provide output uncertainties, and the use of MCS to validate Mainstream GUM.

A model can be viewed as transforming the knowledge of what has been measured or obtained from calibrations, manufacturers' specifications, etc. into knowledge of the required measurement results. It is difficult to consider the errors or uncertainties in this process without this starting point. It is accepted that there may well be difficulties in developing models in some instances. However, their development and recording, even initially in a simple way, provides a basis for review and subsequent improvement.

If, as is permitted by the GUM (Clause G.1.5), other analytical or numerical methods are used, a 95% coverage interval can be obtained, as implied above (using Monte Carlo Simulation, for example), free of these limitations. When these other approaches are employed, there is no concept of a coverage factor.

In this appendix, consideration is given to the relationship of the GUM (input-output) model to classical statistical modelling.

There is much discussion in many quarters of different attitudes to uncertainty evaluation. One attitude involves the formal definition of uncertainty of measurement (GUM Clause 2.2.3 and above) and the other in terms of the concepts of error and true value. These alternative concepts are stated in GUM Clause 2.2.4 as the following characterizations of uncertainty of measurement:

1. a measure of the possible error in the estimated value of the measurand as provided by the result of a measurement;

2. an estimate characterizing the range of values within which the true value of a measurand lies (VIM, first edition, 1984, entry 3.09).

A second edition of VIM has been published [40]. That edition is currently being revised by the JCGM.

The extent and ferocity of discussion on matters such as the existence of a true value almost reaches religious fervour and is such that it provides a major inhibiting influence on progress with uncertainty concepts. This best-practice guide does not subscribe to the view that the attitudes are mutually untenable. Indeed, in some circumstances an analysis can be more straightforward using the concepts of GUM Clause 2.2.3 and in others the use of Clause 2.2.4 confers advantage. It is essential to recognize that, although philosophically the approaches are quite different, whichever concept is adopted an uncertainty component is always evaluated using the same data and related information (GUM Clause 2.2). Indeed, in Clause E.5 of the GUM a detailed comparison of the alternative views of uncertainty clearly reconciles the approaches. In particular, GUM Clause E.5.3 states

> ... it makes no difference *in the calculations* if a standard uncertainty is viewed as a measure of the probability distribution of an input quantity or as a measure of the dispersion of the probability distribution of the error of that quantity.

## B.1 Example to illustrate the two approaches

Consider the measurement of two nominally-identical lengths under suitably-controlled conditions using a steel rule. Suppose there are two contributions to the uncertainty of measurement due to:

1. Imperfection in the manufacture and calibration of the rule,
2. Operator effect in positioning and reading the scale.

Let the true lengths be denoted by $L_1$ and $L_2$. Let the measured lengths be denoted by $\ell_1$ and $\ell_2$. Then the measurements may be modelled by

$$\begin{aligned} \ell_1 &= L_1 + e_0 + e_1, \\ \ell_2 &= L_2 + e_0 + e_2, \end{aligned}$$

where $e_0$ is the imperfection error in the steel rule when measuring lengths close to those of concern, and $e_1$ and $e_2$ are the operator errors in making the measurements. The errors in the measured lengths are therefore

$$\begin{aligned} \ell_1 - L_1 &= e_0 + e_1, \\ \ell_2 - L_2 &= e_0 + e_2. \end{aligned}$$

Make the reasonable assumption that the errors $e_0$, $e_1$ and $e_2$ are independent. Then, if $\sigma_0$ denotes the standard deviation of $e_0$ and $\sigma$ that of $e_1$ and $e_2$,

$$\begin{aligned} \mathrm{var}(\ell_1 - L_1) &= \sigma_0^2 + \sigma^2, \\ \mathrm{var}(\ell_2 - L_2) &= \sigma_0^2 + \sigma^2, \\ \mathrm{cov}(\ell_1 - L_1, \ell_2 - L_2) &= \sigma_0^2. \end{aligned}$$

Suppose that it is required to evaluate the difference in the measured lengths and its uncertainty. From the above equations,

$$\ell_1 - \ell_2 = (L_1 - L_2) + (e_1 - e_2)$$

and hence, since $e_1$ and $e_2$ are independent,

$$\mathrm{var}(\ell_1 - \ell_2) = \mathrm{var}(e_1 - e_2) = \mathrm{var}(e_1) + \mathrm{var}(e_2) = 2\sigma^2.$$

As expected, the uncertainty in the steel rule does not enter this result.
Compare the above with the Mainstream GUM approach:
*Inputs.* $L_1$ and $L_2$.
*Model.* $Y = L_1 - L_2$.
*Input values.* $\ell_1$ and $\ell_2$.
*Input uncertainties.*

$$u(\ell_1) = u(\ell_2) = \left(\sigma_0^2 + \sigma^2\right)^{1/2}, \qquad u(\ell_1, \ell_2) = \sigma_0^2.$$

*Partial derivatives of model* (evaluated at the input values)

$$\partial Y/\partial L_1 = 1, \qquad \partial Y/\partial L_2 = -1.$$

*Output uncertainty using GUM Formula (13)*

$$\begin{aligned} u_c^2(y) &= (\partial Y/\partial L_1)^2\, u^2(\ell_1) + (\partial Y/\partial L_2)^2\, u^2(\ell_2) + 2\,(\partial Y/\partial L_1)\,(\partial Y/\partial L_2)\, u(\ell_1, \ell_2) \\ &= (1)^2(\sigma_0^2 + \sigma^2) + (-1)^2(\sigma_0^2 + \sigma^2) + 2(1)(-1)\sigma_0^2 \\ &= 2\sigma_0^2. \end{aligned}$$

# Appendix C

# The use of software for algebraic differentiation

Sensitivity coefficients can be difficult to determine by hand for models that are complicated. The process by which they are conventionally determined is given in Section 5.5. The partial derivatives required can in principle be obtained using one of the software systems available for determining derivatives automatically by applying the rules of algebraic differentiation.

If such a system is used, care needs to be taken that the mathematical expressions generated are stable with respect to their evaluation at the estimates of the input quantities. For instance, suppose that (part of) a model is

$$Y = (X_1 - C)^4,$$

where $C$ is a specified constant. An automatic system might involve expansions such as Taylor series to generate the partial derivative of $Y$ in the form

$$\frac{\partial Y}{\partial X_1} = 4X_1^3 - 12X_1^2C + 12X_1C^2 - 4C^3, \tag{C.1}$$

and perhaps not contain a facility to generate directly or simplify this expression to the *mathematically* equivalent form

$$\frac{\partial Y}{\partial X_1} = 4(X_1 - C)^3, \tag{C.2}$$

that would typically be obtained *manually*.

Suppose the estimate of $X_1$ is $x_1 = 10.1$ and $C = 9.9$. The value $c_1$ of the resulting sensitivity coefficient is $4(x_1 - C)^3 = 0.032$, correct to two significant figures. Both formulae (C.1) and (C.2) deliver this value to this number of figures. The second, more compact, form is, however, much to be preferred. The reason is that Formula (C.2) suffers negligible loss of numerical precision when it is used to evaluate $c_1$, whereas, in contrast, Formula (C.1) loses figures in forming this value. To see why this the case, consider the contributions to the expression, evaluated and displayed here to a constant number of decimal places (corresponding to seven significant figures in the contribution of greatest

magnitude):

$$\begin{aligned} 4x_1^3 &= 4121.204, \\ -12x_1^2C &= -12118.79, \\ 12x_1C^2 &= 11878.81, \\ -4C^3 &= -3881.196. \end{aligned}$$

The sum of these values constitutes the value of $c_1$. To the numerical accuracy held, this value is 0.028, compared with the correct value of 0.032. The important point is that a value of order $10^{-2}$ has been obtained by the sum of positive and negative values of magnitude up to order $10^4$. Almost inevitably, a loss of some six figures of numerical precision has resulted, as a consequence of *subtractive cancellation*.

For different values of $x_1$ and $C$ or in other situations the loss of figures could be greater or less. The concerning matter is that this loss has resulted from such a simple model. The effects in the case of a sophisticated model or a multi-stage model could well be compounded, with the consequence that there are dangers that the sensitivity coefficients formed in this way will be insufficiently accurate. Therefore, care must be taken in using sensitivity coefficients that are evaluated from the expressions provided by some software for algebraic differentiation. Such a system, if used, should evidently be chosen with care. One criterion in making a choice is whether the system offers comprehensive facilities for carrying out algebraic simplification, thus ameliorating the danger of loss of figures. Even then, some form of validation should be applied to the numerical values so obtained.[1]

Some systems operate in a different way. They "differentiate code" rather than differentiate a formula. Acceptable results can be produced by the user by carefully constructing the code that defines the function.

[1] Numerical analysis issues such as that discussed here will be addressed as part of SS*f*M2.

# Appendix D

# Frequentist and Bayesian attitudes

## D.1 Discussion on Frequentist and Bayesian attitudes

There are strongly-held views concerning whether statistical analysis in general or uncertainty evaluation in particular should be carried out according to Frequentist or Bayesian attitudes.

The Frequentist would assume that the value of the measurand is an unknown constant and that the measurement result is a random variable. The Bayesian would regard the value of the measurand as a random variable having a pdf derived from existing knowledge and the result of measurement as a known quantity [44].

These views can result in such divergent opinions that their discussion, although of considerable philosophical interest, militates against the practical need to provide useful uncertainty evaluations.

The attitude of this guide is to adopt the attitude that there is no "right answer". However, it is only reasonable to adopt the view that appropriate use should be made of the information available. In this regard this guide takes predominantly a Bayesian attitude. A Bayesian would use available knowledge to make judgements, often subjective to some extent, of the pdf's of the model inputs. The practitioner, with support from the expert metrologist as necessary, would also wish to employ previous information, e.g., calibration information or measurements of similar artefacts. Where (some of) the information seems suspect, as a result of common sense, experience or statistical analysis, further information should be sought if it is economical to do so.

In several instances the results that would finally be obtained by Frequentists and Bayesians would be identical or at least similar. Consider the repeated measurement of items manufactured under nominally identical conditions. The Frequentist would analyze the sample of measurements to estimate the "population" mean and standard deviation of the manufactured items, and perhaps other parameters. The Bayesian would devise a prior distribution, based on his knowledge of the manufacturing process. He would "update" the information

contained within it in order to provide hopefully more reliable estimates of the parameters. In a case where there was no usable information available initially, the Bayesian would employ the so-called "non-informative prior". This prior effectively corresponds to the minimal knowledge that in the absence of information any measurement is equally possible. The parameter values so obtained can be identical in this case to those of the Frequentist. Any additional knowledge would help to give a better prior and hopefully a more valid result in that it would depend on available application-dependent information.

## D.2 The Principle of Maximum Entropy

> "... the virtue of wise decisions by taking into account all possibilities, i.e., by not presuming more information than we possess." [43]

The Principle of Maximum Entropy (PME) [42] is a concept that can valuably be employed to enable maximum use to be made of available information, whilst at the same time avoiding the introduction of unacceptable bias in the result obtained. Two internationally respected experts in measurement uncertainty [65] state that predictions based on the results of Bayesian statistics and this principle turned out to be so successful in many fields of science [59], particularly in thermodynamics and quantum mechanics, that experience dictates no reason for not also using the principle in a theory of measurement uncertainty.

Bayesian statistics have been labelled "subjective", but that is the intended nature of the approach. One builds in knowledge based on experience and other information to obtain an improved solution. However, if the same knowledge is available to more than one person, it would be entirely reasonable to ask that they drew the same conclusion. The application of PME was proposed [66] in the field of uncertainty evaluation in order to achieve this objective.

To illustrate the principle, consider a problem [66] in which a single unknown systematic deviation $X_1$ is present in an measurement process. Suppose that all possible values for this deviation lie within an interval $[-L,\ L]$, after the observed measurement result was corrected as carefully as possible for a known constant value. The value supplied for $L$ is a subjective estimate based on known properties of the measurement process, including the model inputs. In principle, the value of $L$ could be improved by aggregating in a suitable manner the estimates of several experienced people. Let $g_1(x_1)$ denote the pdf of $X_1$. Although it is unknown, $g_1(x_1)$ will of course satisfy the normalizing condition

$$\int_{-L}^{L} g_1(x_1)dx_1 = 1. \tag{D.1}$$

Suppose that from the properties of the measurement process it can be asserted that on average the systematic deviation is expected to be zero. A second condition on $g_1(x_1)$ is therefore

$$\int_{-L}^{L} x_1 g_1(x_1)dx_1 = 0. \tag{D.2}$$

Suppose an estimate $u$ of the standard deviation of the possible values for the

systematic deviation is available. Then,

$$\int_{-L}^{L} x_1^2 g_1(x_1) dx_1 = u^2. \tag{D.3}$$

Of course, $u^2 \leq L^2$. Suppose that no further information is available.

There are of course infinitely many pdf's for which the above three conditions hold. However, PME can be used to select a pdf from these. The pdf so obtained will have the property that it will be unbiased in that nothing more is implicitly assumed.

The use [42] of Shannon's theory of information [56] achieves this goal. Any given pdf represents some lack of knowledge of the quantity under consideration. This lack of knowledge can be quantified by a number,

$$S = -\int g_1(x) \log g_1(x) dx,$$

called the (information) entropy [56] of that pdf, The least-biased "probability assignment" is that which maximizes $S$ subject to the conditions (D.1)-(D.3). Any other pdf that satisfies these conditions has a smaller value of $S$, thus introducing a prejudice that might contradict the missing information.

This formulation can fully be treated mathematically [66] and provides the required pdf. The "shape" of the pdf depends on the quotient $u/L$. If $u/L$ is smaller than $1/\sqrt{3}$, the pdf is bell-shaped. If $u/L$ is larger than $1/\sqrt{3}$, the pdf is U-shaped. Between these possibilities, when $u/L = 1/\sqrt{3}$, the pdf is the uniform pdf.

It can also be determined that if no information is available about the standard deviation $u$, and $S$ is maximized with respect to conditions (D.1) and (D.2) only, the resulting pdf is the uniform pdf.

It is evident that the more information there is available the better the required pdf can be estimated. The services of a mathematician or a statistician might be required for this purpose. However, in a competitive environment, or simply when it is required to state the most realistic (and hopefully the smallest) and defensible measurement uncertainties, such a treatment might be considered appropriate.

One approach to a range of such problems might be to categorize the commonest types of problem, such as those above, viz., when

1. Conditions (D.1) and (D.2) hold,
2. Conditions (D.1)-(D.3) hold.

There would other such conditions in other circumstances. "Solutions" to this range of problems could be determined, almost certainly in the form of algorithms that took as input the defining parameters (such as $L$ and $u$ above) and returned the corresponding quantified pdf. This pdf would then form one of the inputs to an MCS solution or to what other method might be applied.[1]

In summary, it is regarded as scientifically flawed to discard credible information, unless it can be shown that doing so will have no influence on the results required to the accuracy needed.

[1] Such considerations would seem valuable for a future SS*f*M programme and thence to feed into a subsequent release of this guide.

In particular, if knowledge of the pdf's of the inputs is available, perhaps deduced as above using PME, these pdf's, which can be regarded as providing prior information in a Bayesian sense, should not simply be replaced by a mean and standard deviation, unless doing so can be shown to have the mentioned negligible effect. If other information is available, such as above, or conditions on the results of measurement or on nuisance parameters,[2] this information should be incorporated in order to render the solution physically more meaningful, and the uncertainties more realistic.

There are some further cases that can be dealt with reasonably straightforwardly:

1. If a series of repeat measurements is taken and no knowledge is available other than the data itself, it can be inferred from the PME that a Gaussian pdf with mean and standard deviation obtained from the measurements should be assigned to the parameter estimated by the measurements.

2. If the measurements are as in 1 above, but are known to contain errors that can be considered as drawn from a Gaussian distribution, it can be inferred from the PME that a Student's $t$ pdf with mean, standard deviation and number of degrees of freedom obtained from the measurements should be assigned.

3. If the situation is as in 2, but that additionally a prior Gaussian pdf is available from historical information, it can be inferred from the PME that taking both sources of information into account, an improved Gaussian pdf can be obtained [49].[3] If $x_P$ is the estimate of the measurand based on prior information (only) and $x_M$ that on the measurements (without including prior information), and $u_P$ and $u_M$ are the corresponding standard deviations, the best estimate [49] of the measurand using both sources of information is
$$x = \left(\frac{1}{1+\gamma^2}\right) x_P + \left(\frac{\gamma^2}{1+\gamma^2}\right) x_M,$$
where
$$\gamma = u_P/u_M,$$
with standard deviation $u(x)$ given by
$$\frac{1}{u^2(x)} = \frac{1}{u_P^2} + \frac{1}{u_M^2}.$$

Clause 4.3.8 of the GUM provides the pdf when limits plus a single measurement are available.

**Example 26** *The application of the Principle of Maximum Entropy to determining the probability density function when lower and upper bounds only are available*

[2] Nuisance parameters are additional variables introduced as part of the modelling process to help build a realistic model. They would not by their nature constitute measurement results, but their values and their uncertainties might be of interest as part of model development or enhancement.

[3] Check whether the pdf is Gaussian or Student's-$t$.

Consider that lower and upper bounds $a$ and $b$ for the input quantity $X_i$ are available. If no further information is available the PME would yield $(a+b)/2$ as the best estimate of $X_i$ and $\{(b-a)/12\}^{1/2}$ as the best estimate of the standard deviation of $X_i$. It would also yield the pdf of $X_i$ as the uniform distribution with limits $a$ and $b$.

∎

**Example 27** *The application of the Principle of Maximum Entropy to determining the probability density function when lower and upper bounds and a single measurement are available*

Suppose that as well as limits $a$ and $b$, an estimate $x_i$ of $X_i$ is available. Unless $x_i = (a+b)/2$, i.e., it lies at the centre of the interval $(a,\ b)$, the pdf would not be uniform as before. Let $\lambda$ be the root of the equation

$$(e^{-\lambda a} - e^{-\lambda b})C(\lambda) - \lambda = 0,$$

where

$$C(\lambda)^{-1} = (x_i - a)e^{-\lambda a} + (b - x_i)e^{-\lambda b}.$$

The PME yields [1, Clause 4.3.8] the pdf of $X_i$ as

$$C(\lambda)e^{-\lambda X_i}.$$

∎

All circumstances should be treated on their merits. Consider a *large* number of repeat measurements. Suppose that the manner in which they are distributed (as seen by a histogram of their values, e.g.) indicates clearly that their behaviour is non-Gaussian, e.g., a strong asymmetry, "long tails" or bi-modality. Then, the data itself, if judged to be representative, is indicating that the blind application of the PME is inappropriate in this circumstance. Since, for large samples, the principle of the bootstrap is appropriate, it can legitimately be applied here.

# Appendix E

# Nonlinear sensitivity coefficients

Nonlinear sensitivity coefficients offer a possible counterpart of the sensitivity coefficients that are a major component of the Mainstream GUM approach. They apply quite generally, i.e., to any model and any input pdf's, regardless of the extent of model nonlinearity or complexity or the nature of the pdf's.

With a linear model the sensitivity coefficients "reproduce" linear effects. For a nonlinear model, the sensitivity coefficients provide first-order information. With MCS there is no immediate counterpart of a sensitivity coefficient since MCS operates in terms of the actual nonlinear model rather than a linearized counterpart. Therefore, those practitioners accustomed to the "GUM culture" may find the absence of sensitivity coefficients disconcerting. There is no counterpart of a (constant) coefficient in the general setting. It is possible and very straightforward, however, to adapt MCS such that it provides information that in a sense constitutes a nonlinear counterpart of a sensitivity coefficient. Consider holding all inputs but one, that of concern, at their nominal values. In this setting the model effectively becomes one having a single input quantity. Sample values from this input and provide the pdf of the output quantity with respect to this input quantity. This pdf provides a distributional statement of uncertainty. In the case of complicated models this approach is already used by many metrologists as a practical alternative to the tedious analysis required to provide sensitivity coefficients [36].

The quotient of the standard deviation of this pdf to that of the relevant input quantity can be taken as a "nonlinear" sensitivity coefficient.