TECHNICAL REPORT

ISO/TR 13587

First edition
2012-07-15

Three statistical approaches for the assessment and interpretation of measurement uncertainty

Trois approches statistiques pour l'évaluation et l'interprétation de l'incertitude de mesure

Reference number
ISO/TR 13587:2012(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In exceptional circumstances, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide by a simple majority vote of its participating members to publish a Technical Report. A Technical Report is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TR 13587:2012 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods*, Subcommittee SC 6, *Measurement methods and results*.

This Technical Report is primarily based on Reference [10].

Introduction

The adoption of ISO/IEC Guide 98-3 (GUM) [1] has led to an increasing recognition of the need to include uncertainty statements in measurement results. Laboratory accreditation based on International Standards like ISO 17025 [2] has accelerated this process. Recognizing that uncertainty statements are required for effective decision-making, metrologists in laboratories of all types, from National Metrology Institutes to commercial calibration laboratories, are exerting considerable effort on the development of appropriate uncertainty evaluations for different types of measurement using methods given in the GUM.

Some of the strengths of the procedures outlined and popularized in the GUM are its standardized approach to uncertainty evaluation, its accommodation of sources of uncertainty that are evaluated either statistically (Type A) or non-statistically (Type B), and its emphasis on reporting all sources of uncertainty considered. The main approach to uncertainty propagation in the GUM, based on linear approximation of the measurement function, is generally simple to carry out and in many practical situations gives results that are similar to those obtained more formally. In short, since its adoption, the GUM has sparked a revolution in uncertainty evaluation.

Of course, there will always be more work needed to improve the evaluation of uncertainty in particular applications and to extend it to cover additional areas. Among such other work, the Joint Committee for Guides in Metrology (JCGM), responsible for the GUM since the year 2000, has completed Supplement 1 to the GUM, namely, “Propagation of distributions using a Monte Carlo method” (referred to as GUMS1) [3]. The JCGM is developing other supplements to the GUM on topics such as modelling and models with any number of output quantities.

Because it should apply to the widest possible set of measurement problems, the definition of measurement uncertainty in ISO/IEC Guide 99:2007 [4] as a “non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used” cannot reasonably be given at more than a relatively conceptual level. As a result, defining and understanding the appropriate roles of different statistical quantities in uncertainty evaluation, even for relatively well-understood measurement applications, is a topic of particular interest to both statisticians and metrologists.

Earlier investigations have approached these topics from a metrological point of view, some authors focusing on characterizing statistical properties of the procedures given in the GUM. Reference [5] shows that these procedures are not strictly consistent with either a Bayesian or frequentist interpretation. Reference [6] proposes some minor modifications to the GUM procedures that bring the results into closer agreement with a Bayesian interpretation in some situations. Reference [7] discusses the relationship between procedures for uncertainty evaluation proposed in GUMS1 and the results of a Bayesian analysis for a particular class of models. Reference [8] also discusses different possible probabilistic interpretations of coverage intervals and recommends approximating the posterior distributions for this class of Bayesian analyses by probability distributions from the Pearson family of distributions.

Reference [9] compares frequentist (“conventional”) and Bayesian approaches to uncertainty evaluation. However, the study is limited to measurement systems for which all sources of uncertainty can be evaluated using Type A methods. In contrast, measurement systems with sources of uncertainty evaluated using both Type A and Type B methods are treated in this Technical Report and are illustrated using several examples, including one of the examples from Annex H of the GUM.

Statisticians have historically placed strong emphasis on using methods for uncertainty evaluation that have probabilistic justification or interpretation. Through their work, often outside metrology, several different approaches for statistical inference relevant to uncertainty evaluation have been developed. This Technical Report presents some of those approaches to uncertainty evaluation from a statistical point of view and relates them to the methods that are currently being used in metrology or are being developed within the metrology community. The particular statistical approaches under which different methods for uncertainty evaluation will be described are the frequentist, Bayesian, and fiducial approaches, which are discussed further after outlining the notational conventions needed to distinguish different types of quantities.

Three statistical approaches for the assessment and interpretation of measurement uncertainty

1 Scope

This Technical Report is concerned with three basic statistical approaches for the evaluation and interpretation of measurement uncertainty: the frequentist approach including bootstrap uncertainty intervals, the Bayesian approach, and fiducial inference. The common feature of these approaches is a clearly delineated probabilistic interpretation or justification for the resulting uncertainty intervals. For each approach, the basic method is described and the fundamental underlying assumptions and the probabilistic interpretation of the resulting uncertainty are discussed. Each of the approaches is illustrated using two examples, including an example from ISO/IEC Guide 98-3 (*Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)*). In addition, this document also includes a discussion of the relationship between the methods proposed in the GUM Supplement 1 and these three statistical approaches.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 3534-1:2006, *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*

ISO 3534-2:2006, *Statistics — Vocabulary and symbols — Part 2: Applied statistics*

ISO/IEC Guide 98-3:2008, *Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)*

ISO/IEC Guide 98-3:2008/Suppl 1:2008, *Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995) — Supplement 1: Propagation of distributions using a Monte Carlo method*

3 Terms and definitions

For the purposes of this document, the terms and definitions in ISO 3534-1, ISO 3534-2 and the following apply.

3.1
empirical distribution function
empirical cumulative distribution function
distribution function that assigns probability $1/n$ to each of the n items in a random sample, i.e., the empirical distribution function is a step function defined by

$$F_n(x) = \frac{\left|\{x_i \le x\}\right|}{n},$$

where $\{x_1,...,x_n\}$ is the sample and $|A|$ is the number of elements in the set A.

3.2
Bayesian sensitivity analysis
study of the effect of the choices of prior distributions for the parameters of the statistical model on the posterior distribution of the measurand

3.3
sufficient statistic
function of a random sample $X_1,...,X_n$ from a probability density function with parameter θ for which the conditional distribution of $X_1,...,X_n$ given this function does not depend on θ

NOTE A sufficient statistic contains as much information about θ as $X_1,...,X_n$.

3.4
observation model
mathematical relation between a set of measurements (indications), the measurand, and the associated random measurement errors

3.5
structural equation
statistical model relating the observable random variable to the unknown parameters and an unobservable random variable whose distribution is known and free of unknown parameters

3.6
non-central chi-squared distribution
probability distribution that generalizes the typical (or central) chi-squared distribution

NOTE 1 For k independent, normally distributed random variables X_i with mean μ_i and variance σ_i^2, the random variable $X = \sum_{i=1}^{k} (X_i/\sigma_i)^2$ is non-central chi-squared distributed. The non-central chi-squared distribution has two parameters: k, the degrees of freedom (i.e., the number of X_i), and λ, which is related to the means of the random variables X_i by $\lambda = \sum_{i=1}^{k} (\mu_i/\sigma_i)^2$ and called the non-centrality parameter.

NOTE 2 The corresponding probability density function is expressed as a mixture of central χ^2 probability density functions as given by

$$g_X(\xi) = \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^i}{i!} g_{Y_{k+2i}}(\xi) = \frac{e^{\frac{-(\xi+\lambda)}{2}}}{2^{\frac{k}{2}}} \sum_{i=0}^{\infty} \frac{\xi^{\frac{k}{2}+i-1}\lambda^i}{\Gamma\left(\frac{k}{2}+i\right)2^{2i}i!},$$

where Y_q is distributed as chi-squared with q degrees of freedom.

4 Symbols (and abbreviated terms)

In 4.1.1 of the GUM, it is stated that Latin letters are used to represent both physical quantities to be determined by measurement (i.e., measurands in GUM terminology) as well as random variables that may take different observed values of a physical quantity. This use of the same symbols, whose different meanings are only indicated by context, can be difficult to interpret and sometimes leads to unnecessary ambiguities or misunderstandings. To mitigate this potential source of confusion, the more traditional notation often used in the statistical literature is employed in this Technical Report. In this notation, Greek letters are used to represent parameters in a statistical model (e.g., measurands), which can be either random variables or

constants depending on the statistical approach being used and nature of the model. Upper-case Latin letters are used to represent random variables that can take different values of an observable quantity (e.g., potential measured values), and lower-case Latin letters to represent specific observed values of a quantity (e.g., specific measured values). Since additional notation may be required to denote other physical, mathematical, or statistical concepts, there will still always be some possibility for ambiguity[1)]. In those cases the context clarifies the appropriate interpretation.

5 The problem addressed

5.1 The concern in this Technical Report is with a measurement model in which $\mu_1,...,\mu_p$ are input quantities and θ is the output quantity:

$$\theta = f\left(\mu_1,...,\mu_p\right), \tag{1}$$

where f is known as the measurement function. The function f is specified mathematically or as a calculation procedure. In the GUM (4.1, NOTE 1), the same functional relationship is given as

$$Y = f\left(X_1,...,X_p\right) \tag{2}$$

which cannot be easily distinguished from the measurement function evaluated at the values of the corresponding random variables for each observed input.

Using the procedure recommended in the GUM, the p unknown quantities $\mu_1,...,\mu_p$ are estimated by values $x_1,...,x_p$ obtained from physical measurement or from other sources. Their associated standard uncertainties are also obtained from the relevant data by statistical methods or from probability density functions based on expert knowledge that characterize the variables. The GUM (also see 4.5 in Reference [11]) recommends that the same measurement model that relates the measurand θ to the input quantities $\mu_1,...,\mu_p$ be used to calculate y from $x_1,...,x_p$. Thus, the measured value (or, in statistical nomenclature, the estimate) y of θ is obtained as

$$y = f(x_1,...,x_p), \tag{3}$$

that is, the evaluated Y , $y = f(x_1,...x_p)$, is taken to be the measured value of θ. The estimates y , $x_1,...,x_p$ are realizations of $Y, X_1,..., X_p$, respectively.

5.2 In this Technical Report, three statistical approaches are each used to provide (a) a best estimate y of θ, (b) the associated standard uncertainty $u(y)$, and (c) a confidence interval or coverage interval for θ for a prescribed coverage probability (often taken as 95 %).

5.3 When discussing standard uncertainties, distinction is made between evaluated standard uncertainties associated with estimates of various quantities and their corresponding theoretical values. Accordingly, notation such as σ_μ or σ_X will denote theoretical standard uncertainties and notation such as S_X and s_x will denote an evaluated standard uncertainty before and after being observed, respectively.

1) For example, not all quantities represented by Greek letters in a statistical model must be parameters of the model. One common example of this type of quantity is the set of unobservable quantities that represent the random measurement errors found in most statistical models (i.e., the ε_i in the model $Y_i = \mu + \varepsilon_i$).

6 Statistical approaches

6.1 Frequentist approach

6.1.1 The first statistical approach to be considered, in which uncertainty can be evaluated probabilistically, is frequentist. The frequentist approach is sometimes referred to as “classical” or “conventional”. However, due to the nature of uncertainty in metrology, these familiar methods must often be adapted to obtain frequentist uncertainty intervals under realistic conditions.

6.1.2 In the frequentist approach, the input quantities $\mu_1,...,\mu_p$ in the measurement model (1) and the output quantity θ are regarded as unknown constants. Then, data related to each input parameter, μ_i, is obtained and used to estimate the value of θ based on the measurement model or the corresponding statistical models. Finally, confidence intervals for θ, for a specified level of confidence, are obtained using one of several mathematical principles or procedures, for example, least-squares, maximum likelihood, or the bootstrap.

6.1.3 Because θ is treated as a constant, a probabilistic statement associated with a confidence interval for θ is not a direct probability statement about its value. Instead, it is a probability statement about how frequently the procedure used to obtain the uncertainty interval for the measurand would encompass the value of θ with repeated use. “Repeated use” means that the uncertainty evaluation is replicated many times using different data drawn from the same distributions. Traditional frequentist uncertainty intervals provide a probability statement about the long-run properties of the procedure used to construct the interval under the particular set of conditions assumed to apply to the measurement process.

6.1.4 In most practical metrological settings, on the other hand, uncertainty intervals are to account for the uncertainty associated with estimates of quantities obtained using measured values (observed data) and also the uncertainty associated with estimates of quantities based on expert knowledge. To obtain an uncertainty interval analogous to a confidence interval, the quantities that are not based on measured values are treated as random variables with probability distributions for their values while those quantities whose values can be estimated using statistical data are treated as unknown constants.

6.1.5 Traditional frequentist procedures for the construction of confidence intervals are then to be modified to attain the specified confidence level after averaging over the potential values of the quantities assessed using expert judgment [5]. Such modified coverage intervals provide long-run probability statements about the procedure used to obtain the interval given probability distributions for the quantities that have not been measured, just as traditional confidence intervals do when all parameters are treated as constants.

6.1.6 Table 1 summarizes interpretations of the frequentist, Bayesian and fiducial approaches to uncertainty evaluation.

Table 1 — Interpretations of the approaches to uncertainty evaluation

Approach	Characterization of quantities in measurement model $\theta = f(\mu_1,...,\mu_p)$	Uncertainty interval for output quantity θ	Note
Frequentist	θ and the μ_i all unknown constants	Long-run occurrence frequency that interval contains θ	Classical frequentist approach extended to integrate over uncertainties that are not statistically evaluated
Bayesian	θ and the μ_i are random variables. Their probability distributions represent beliefs about the values of the input and output quantities	Coverage interval containing θ based on a posterior distribution for θ	Possible non-uniqueness of interval due to the choice of priors
Fiducial	μ_i regarded as random variables whose distributions are obtained from assumptions on observed data used to estimate μ_i and expert knowledge about μ_i	Coverage interval containing θ based on a fiducial distribution for θ	Non-uniqueness due to the choice of the structural equation

6.2 Bayesian approach

The second approach is called the Bayesian approach. It is named after the fundamental theorem on which it is based, which was proved by the Reverend Thomas Bayes in the mid-1700s [12]. In this approach, knowledge about the quantities in measurement model (1) in Clause 5 is modelled as a set of random variables that follow a joint probability distribution for $\mu_1,...,\mu_p$ and θ. Bayes' theorem then allows these probability distributions to be updated based on the observed data (also modelled using probability distributions) and the interrelationships of the parameters defined by the function f or equivalent statistical models. Then, a probability distribution is obtained that describes knowledge of θ given the observed data. Uncertainty intervals that contain θ with any specified probability can then be obtained from this distribution. Because knowledge of the parameter values is described by probability distributions, Bayesian methods provide direct probabilistic statements about the value of θ and the other parameters, using a definition of probability as a measure of belief.

6.3 Fiducial approach

6.3.1 The fiducial approach was developed by R.A. Fisher [13] in the 1930s. In this approach, a probability distribution, called the fiducial distribution, for θ conditional on the data is obtained based on the interrelationship of θ and the μ_i described by f and the distributional assumptions about the data used to estimate the μ_i. Once obtained, the fiducial distribution for θ can be used to obtain uncertainty intervals that contain θ with any specified probability.

6.3.2 The argument that justifies the process used to obtain the fiducial distribution is illustrated using a simple example. Suppose the values taken by a quantity Y can be described by the equation $Y = \mu + Z$, where μ is the measurand and Z is a quantity characterized by a standard normal random variable. If y is a realized value of Y corresponding to a realized value z of Z, then $\mu = y - z$. Despite Z not being observable, knowledge of the distribution from which z was generated enables a set of plausible values of μ to be determined. The probability distribution for Z can be used to infer the probability distribution for μ. The

process of transferring the relationship $\mu = y - z$ to the relation $\mu = y - Z$ is what constitutes the fiducial argument. The fiducial distribution for μ is the probability distribution for the random variable $y - Z$ with y fixed.

6.4 Discussion

When describing the different methods for uncertainty evaluation under each of these statistical approaches, their fundamental underlying assumptions, incorporation of uncertainties obtained using Type A or Type B evaluation, and the probabilistic interpretation of the resulting uncertainty evaluations will be discussed. A description of how the methods used in the GUM relate to the frequentist, Bayesian, or fiducial results will also be given.

7 Examples

7.1 General

Two examples are given to illustrate the approaches. Example 1 is concerned with a physical quantity that is to be corrected for background interference. Table 2 gives the notation used and Subclauses 7.2 to 7.4 define variants of this evaluation problem. Example 2 is the calibration of the length of a gauge block taken from Annex H.1 of the GUM. Because it is more complicated, it is considered in Clause 11, after the three methods for uncertainty evaluation are discussed and illustrated using Example 1.

In later clauses, the three approaches will be applied to these examples.

NOTE The units of the quantities involved are not given when they are immaterial for the example.

Table 2 — Notation for Example 1

Quantity	Symbol
Physical quantity of interest (the measurand)	θ
Quantity detected by measurement method when measuring background (i.e., expected value of B) (Background interference)	β
Quantity detected by measurement method when measuring the physical quantity of interest (i.e., expected value of Y)	$\gamma = \theta + \beta$
Standard deviation of measurement method when measuring the physical quantity of interest (i.e., standard deviation of Y)	σ_Y
Standard deviation of measurement method when measuring background (i.e., standard deviation of B)	σ_B

7.2 Example 1a

Five measured values, obtained independently, of signal plus background are observed. Each measured value is assumed to be a realization of a random variable, Y , having a Gaussian distribution with mean $\gamma = \theta + \beta$ and standard deviation σ_Y . The measured values, y , of the signal plus background are

3,738, 3,442, 2,994, 3,637, 3,874.

This data has a sample mean of $\bar{y} = 3,537$ and a sample standard deviation of $s_y = 0,342$.

Similarly, five measured values, obtained independently, of the background are obtained. These measured values are assumed to be realizations of a random variable, B, having a Gaussian distribution with mean β and standard deviation σ_B. The observed values, b, of the background are

1,410, 1,085, 1,306, 1,137, 1,200.

Because there are measured values for each quantity that is a source of uncertainty, Example 1a has a straightforward statistical interpretation for each approach.

7.3 Example 1b

Example 1b is identical to Example 1a with the exception that the assessment of the background is based on expert knowledge or past experience, rather than on fresh experimental data. In this case, the background β is believed to follow a uniform (or rectangular) distribution with endpoints 1,126 and 1,329. Because expert judgment is applied, the uncertainty associated with a value of the background will be obtained using a Type B evaluation. Thus, Example 1b can be considered closer than Example 1a to a real measurement situation.

7.4 Example 1c

Example 1c is identical to Example 1b except that the signal θ is closer to the background. The data observed for the signal plus background in this case are

1,340, 1,078, 1,114, 1,256, 1,192.

With the signal just above the background, Example 1c illustrates how physical constraints can be incorporated in the evaluation of uncertainty for each approach.

8 Frequentist approach to uncertainty evaluation

8.1 Basic method

8.1.1 In the frequentist context, parameters are unknown constants. Following the convention to denote random variables by upper case letters and observed values of random variables by lower case letters, a confidence interval can be obtained from a *pivotal quantity* for θ, i.e., a function $W(Y,\theta)$ of the (possibly multivariate) data Y and the parameter θ, whose probability distribution is parameter-free (provided such a distribution can be determined.) Then, a $100(1-\alpha)$ % confidence interval for θ can be determined by calculating lower and upper percentiles ℓ_α and u_α to satisfy $P_\theta(\ell_\alpha \le W(Y,\theta) \le u_\alpha) = 1-\alpha$.

8.1.2 For example, let $Y = (Y_1,...,Y_n)$ be random variables, distributed as $N(\mu,\sigma^2)$, with the further random variable $\bar{Y} = \sum_{i=1}^{n} Y_i / n$. If the parameter of interest is μ, then for known σ, $Z = \dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$

is a pivotal quantity. The frequentist confidence interval for μ is

$$\bar{Y} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \tag{4}$$

where z_β is the $100\,\beta$ percentile of the standardized normal distribution.

If σ is not known, it can be estimated by the sample standard deviation

$$S = \sqrt{\frac{\sum_{j=1}^{n}\left(Y_j - \bar{Y}\right)^2}{n-1}}$$

Then, the (exact) pivotal quantity for μ is obtained by replacing σ in interval (4) by S:

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t(n-1). \tag{5}$$

Thus, a 100($1-\alpha$) % confidence interval for μ based on the Student's *t*-distribution is

$$\bar{Y} \pm \frac{S}{\sqrt{n}} t_{n-1,1-\alpha/2},$$

where $t_{n-1,\beta}$ is the 100β percentile of the *t*-distribution with $n-1$ degrees of freedom.

8.1.3 Instead of *exact* pivotal quantities, which exist only in simple situations, approximate pivotal quantities are commonly employed in applications. For large samples, the central limit theorem can be invoked to obtain approximate confidence intervals based on the normal distribution.

8.1.4 Further methods of obtaining confidence intervals (inverting a test statistic, pivoting a continuous cumulative distribution function, ordering the discrete sample values according to their probabilities, etc.) are discussed in Reference [14]. Some of them are mentioned in Example 1. A computer-intensive method, called the bootstrap, also can be used to construct a confidence interval for pivotal quantities that have unknown distributions. The bootstrap procedure is discussed in 8.2.

8.1.5 Although not explicitly given a frequentist justification from fundamental scientific considerations, the procedures recommended in the GUM can be used to obtain an approximate confidence interval for the measurand. Such confidence intervals are based on an approximate pivotal quantity with an assumed *t*-distribution obtainable from the measurement model (1). Under this procedure, the unknown quantities $\mu_1,...,\mu_p$ are estimated by values $x_1,\cdots,x_p$ obtained from physical measurement or from other sources. Some of the values x_i might be sample means or other functions of data designed to estimate the quantities $\mu_i, i=1,...,m$. Their associated standard uncertainties $u(x_i)$ are also evaluated from the data by statistical methods, typically using the sample standard deviation or using robust rank-based procedures. Such methods are known as Type A evaluations of uncertainty. The degrees of freedom ν_i associated with $u(x_i)$ is determined from the sample size used to estimate μ_i.

8.1.6 Since physical measurements might not always be possible or feasible for some of the μ_i, estimates x_i of μ_i for some i, say $i=m+1,...,p$, are obtained by subjective (or potentially subjective) evaluations, and used together with x_i, for $i=1,...,p$, obtained from Type A evaluations of uncertainty. Thus, non-statistical types of information are used to estimate $\mu_{m+1},...,\mu_p$ using Type B evaluations of uncertainty, including scientific judgment, manufacturer's specifications, or other indirectly related or incompletely specified information.

NOTE Sometimes uncertainties are obtained by both Type A and Type B evaluations of uncertainty.

8.1.7 The GUM recommends that the same measurement model relating the measurand θ to the input quantities $\mu_1,...,\mu_p$ be used to calculate y from $x_1,...,x_p$. Thus, the measured value (or the estimate) y of θ is obtained as

$$y = f(x_1,...,x_m,x_{m+1},...,x_p),$$

that is, the evaluated Y, $y = f(x_1,...x_p)$, is taken to be the measured value of θ.

8.1.8 In the GUM, the law of propagation of uncertainty is used to evaluate the standard uncertainty, $u(y)$. associated with y. The standard uncertainties $u(x_1),...,u(x_p)$ associated with the values $\mathbf{x} = (x_1,...,x_p)$ are used in the Taylor series expansion of the function $f(x_1,...x_p)$ at $\mu_1,...,\mu_p$, whose terms up to first order are

$$f(x_1,...,x_p) \approx f(\mu_1,...,\mu_p) + \sum_{i=1}^{p} c_i(x_i - \mu_i). \tag{6}$$

Denoting $(\mu_1,...,\mu_p)$ by $\boldsymbol{\mu}$, the partial derivatives

$$c_i = \left.\frac{\partial f}{\partial \mu_i}\right|_{\boldsymbol{\mu}=\mathbf{x}}$$

are called sensitivity coefficients. Applying the law of propagation of uncertainty in the GUM gives the approximate standard uncertainty associated with y:

$$u(y) \approx \sqrt{\sum_{i=1}^{p} c_i^2 u^2(x_i) + 2\sum_{i<j} c_i c_j u(x_i, x_j)}, \tag{7}$$

where $u(x_i, x_j)$ is the covariance between X_i and X_j.

8.1.9 To evaluate the standard uncertainty $u(y)$, the GUM uses the effective degrees of freedom ν_{eff} computed from the Welch-Satterthwaite formula,

$$\nu_{\text{eff}} = \frac{u^4(y)}{\sum_{i=1}^{p} \frac{c_i^4 u^4(x_i)}{\nu(x_i)}}. \tag{8}$$

NOTE Reference [15] discusses a counter-intuitive property according to which in interlaboratory studies a confidence interval based on the Welch-Satterthwaite approximation may be shorter for a between-laboratory difference than for one of its components.

8.1.10 Finally, in order to construct a confidence interval for θ, the approximate pivotal quantity,

$$W(y,\theta) = \frac{y - \theta}{u(y)} \tag{9}$$

is employed. According to the GUM,

$$W(Y,\theta) \sim t(\nu_{\text{eff}}), \tag{10}$$

that is, $W(Y,\theta)$ is an approximately pivotal quantity having a *t*-distribution with ν_{eff} degrees of freedom. The $100(1-\alpha)$ % confidence interval

$$y \pm u(y) t_{\nu_{\text{eff}},1-\alpha/2}, \tag{11}$$

for θ can then be recommended as the 100(1 – α) % uncertainty interval for θ. The half-width $t_{\nu_{\text{eff}},1-\alpha/2}u(y)$ of this interval is known as the expanded uncertainty associated with y.

8.1.11 This recommendation agrees with standard statistical practice when all uncertainties are determined using Type A evaluation, in which case the most commonly used statistical estimate for a particular input quantity μ is the sample mean of n observed values. The traditional method for summarizing data to obtain the Type A standard uncertainty of this estimator is $S/\sqrt{n}$ with $n-1$ degrees of freedom. This is based on the fact that $(n-1)S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom. This method applies to more general statistics of the form $Y = G(X_1,...,X_p)$, where estimators $X_i, i = 1,...,p$ obey the central limit theorem. Indeed in this situation, the standard deviation of Y can be approximated by Expression (7) with $u(x_i,x_j)$ replaced by $\mathrm{Cov}(X_i,X_j)$.

The GUM method presents the collective wisdom of many metrologists, but is restricted by assumptions of

— local linearity of the function f : ideally the sensitivity coefficients should not vary much and not vanish;

— normality of the probability distribution of point estimators $Y = f(X_1,...,X_p)$: may not hold even approximately for small samples;

— validity of the Welch-Satterthwaite Formula (8): it may not work well when the input quantities are mutually dependent, the input quantities are not normally distributed, and the standard uncertainties are dissimilar (degrees of freedom for distributions unrelated to the chi-squared law are difficult to interpret, indeed, they are not used in statistical theory).

8.1.12 To motivate Expression (7) in the frequentist setting, the concepts of statistical decision theory can be employed and the variance (squared standard uncertainty) $u^2(y)$ interpreted as the mean squared error of the statistical estimator of $f(x_1,x_2,...,x_p)$. These steps can be taken provided that the quantities whose uncertainties are determined using a Type B evaluation, namely, $x_{m+1},...,x_p$, are eliminated by integrating over their distributions. See Reference [5]. If f "is sufficiently close to being linear", Expression (7) provides the first order approximation of the mean squared error.

8.1.13 The discussion in Example 1 gives another customary frequentist procedure for obtaining confidence intervals.

8.2 Bootstrap uncertainty intervals

8.2.1 Bootstrapping is a resampling strategy [16] for estimating distribution parameters such as variance and determining confidence intervals for parameters when the form of the underlying distribution is unknown. The key idea for the bootstrap method is that the relation between the cumulative probability distribution (CDF) F for Y and a sample from F is similar to the relation between an estimated CDF $\hat{F}$, which may be not the empirical distribution generated by the sample and a second sample drawn from $\hat{F}$. When F is not available, draws cannot be made from it, but modern computers allow a large number of draws to be made from $\hat{F}$ So, one uses the primary sample to form an approximation $\hat{F}$ of F, and then calculates the sampling distribution of the parameter estimate based on $\hat{F}$. This calculation is carried out by drawing many secondary samples and forming the estimate (or a function of the estimate) for each of the secondary sample. If $\hat{F}$ is a good approximation to F, then H, the sampling distribution of the estimate based on $\hat{F}$, is generally a good approximation to the sampling distribution for the estimate based on F. H is commonly called the bootstrap distribution of the parameter.

8.2.2 There are two types of bootstrap procedures useful, respectively, for non-parametric and parametric inference. The non-parametric bootstrap relies on the consideration of the empirical distribution $\hat{F}$ generated by the primary sample from F. In the parametric bootstrap setting, the probability distribution F is a member of some prescribed parameter family and $\hat{F}$ is obtained by estimating the parameter(s) from the data.

NOTE Since in typical metrological problems, data sets are not large enough to ensure the validity of the non-parametric bootstrap approach, that approach is not considered here.

8.2.3 The key assumption used in constructing the GUM confidence interval is (10), which may not hold approximately even for simple problems. However, the bootstrap enables confidence intervals to be obtained that do not involve making assumptions like (10). One way to obtain such intervals is the "bootstrap-*t*" approach. This procedure generates an empirical distribution for the approximate pivotal quantity $W(Y,\theta)$ [to replace the *t*-distribution in (10)]. When (10) is correct, the bootstrap-*t* distribution will reproduce the *t*-distribution. The empirical bootstrap-*t* distribution is then used to construct a confidence interval in exactly the same way that the *t*-distribution is used in constructing (11).

For the relationship between bootstrapping and the methods proposed in GUMS1, see 12.2.

8.2.4 An outline of the generation of a bootstrap sample is as follows. Assume that x_1 and $u(x_1)$ are the mean and standard deviation for the random variable X_1, which is assumed to follow a probability distribution in some prescribed parameter family. Here for illustration, a Gaussian distribution is used:

a) x_1 and $u(x_1)$ are the estimated mean and standard deviation of a random sample of size k from a Gaussian distribution.

b) From $N(x_1, u^2(x_1))$, generate a sample with sample size k, that is, $\{x^*_{1,1},...,x^*_{1,k}\}$.

c) From $\{x^*_{1,1},...,x^*_{1,k}\}$, calculate the sample mean x^*_1 and the sample standard error $u(x^*_1)$.

$\{x^*_1, u(x^*_1)\}$ is a bootstrap sample of X_1. Similarly, for a given number B, B bootstrap samples can be generated for any variable.

8.2.5 Just as the GUM takes $(x_i, u(x_i))$, for $i = 1,..., p$, as its input to produce y, $u(y)$, and $W(Y,\theta)$, the bootstrap samples $\{x^*_i, u(x^*_i)\}$, $i = 1,..., p$ (see 8.2.4) can be taken as input, to compute $y^*, u(y^*)$, and

$$W^* = W(y^*, y) = \frac{y^* - y}{u(y^*)}. \tag{12}$$

8.2.6 To obtain a bootstrap distribution for $W(Y,\theta)$, for a suitably large B, say 100 000, generate B bootstrap samples $\{x^*_i(b), u(x^*_i(b))\}$, $i = 1,..., p$, and for each compute $W^*(b)$, $b = 1,..., B$. The 100 α th percentile of the bootstrap-*t* distribution of $W(Y,\theta)$ is then approximated by the value $\hat{t}_\alpha$ such that

$$|\{W^*(b) \le \hat{t}_\alpha\}|/B = \alpha,$$

where $|A|$ is the number of elements in the set A. Finally, the 100$(1-\alpha)$ % bootstrap-*t* confidence interval is

$$(y - \hat{t}_{1-\alpha/2} \cdot u(y) \;\; y + \hat{t}_{\alpha/2} \cdot u(y)). \tag{13}$$

The Student-*t* percentiles are symmetric about zero, and as a consequence, (11) must always be symmetric about y. In contrast, the bootstrap-*t* percentiles used in (13) can be asymmetric about zero, leading to an asymmetric uncertainty interval about y, which may provide a more accurate description of the physical situation in some applications. The details of this process in constructing a 95 % uncertainty interval are shown in the following algorithm.

a) For $i = 1,..., p$, using the given distributions for the X_i, generate B bootstrap samples $(x^*_i(1), u(x^*_i(1))),..., (x^*_i(B), u(x^*_i(B)))$.

b) For each bootstrap sample $(x^*_i(b), u(x^*_i(b)))$, $i = 1,..., p$ and $b = 1,..., B$, compute $y^*(b)$, $u(y^*(b))$, and $W^*(b) = (y^*(b) - y)/u(y^*(b))$ following the GUM.

c) Estimate the 100α th percentile of the bootstrap-*t* distribution of $W(Y,\theta)$ by the value $\hat{t}_\alpha$ such that $|\{W^*(b) \le \hat{t}_\alpha\}|/B = \alpha$.

d) Form the 95 % bootstrap-*t* confidence interval $(y - \hat{t}_{0,975} \cdot u(y)\ \ y + \hat{t}_{0,025} \cdot u(y))$.

8.2.7 Bootstrap samples can also be used to replace $u(y)$ by estimating the standard deviation of Y, when the Taylor approximation (6) is deemed inappropriate. To do so, for $i = 1,..., p$ and $b = 1,..., B$, only input estimates $x_i^*(b)$ are generated. For each bootstrap sample, $y^*(b) = f(x_1^*(b),..., x_p^*(b))$ is evaluated. The bootstrap estimate of the standard uncertainty associated with y is the sample standard deviation of the B replicates:

$$u_c(y) = \sqrt{\sum_{b=1}^{B} [y^*(b) - y^*(\cdot)]^2/(B-1)}\ , \qquad y^*(\cdot) = \sum_{b=1}^{B} y^*(b)/B\ .$$

8.2.8 Finally, when the Taylor approximation may be inappropriate and there is significant asymmetry in the underlying distribution for Y, a nested bootstrap of B_1 × B_2 bootstrap samples can be carried out to construct a bootstrap-*t* interval using the bootstrap standard deviation estimator. B_1 bootstrap samples of input estimates and the corresponding y^* are generated. For each bootstrap sample, $u_c(y^*)$ is computed by B_2 second-level bootstrap samples, and

$$\frac{y^* - y}{u(y^*)}$$

is evaluated. The collection of B_1 such ratios is then used to estimate percentiles of $W(Y,\theta)$, which leads to the construction of a bootstrap-*t* interval as in (13). An algorithm to construct a 95 % uncertainty interval using the nested bootstrap is as follows.

a) For $i = 1,..., p$, using the distribution for the X_i, generate B_1 first-level bootstrap samples $x_i^*(1),..., x_i^*(B_1)$.

b) For each first-level bootstrap sample $x_i^*(b_1)$, $i = 1,..., p$, $b_1 = 1,..., B_1$ compute $y^*(b_1) = f(x_1^*(b_1),..., x_p^*(b_1))$, and $W^*(b_1) = (y^*(b_1) - y)/u(y^*(b_1))$, where $u(y^*(b_1))$ is determined by a second-level bootstrap using the following algorithm:

   1. For $i = 1,..., p$, using the distribution for μ_i, generate B_2 second-level bootstrap samples $x_i^*(1),..., x_i^*(B_2)$.

   2. For each second-level bootstrap sample, evaluate $y^*(b_2) = f(x_1^*(b_2),..., x_p^*(b_2))$.

   3. Form the bootstrap estimate of the standard uncertainty of $y^*(b_1)$ as the sample standard deviation

$$u(y^*(b_1)) = \sqrt{\sum_{b_2=1}^{B_2} [y^*(b_2) - y^*(\cdot)]^2/(B_2 - 1)}$$

   of the B_2 replicates, where $y^*(\cdot) = \sum_{b_2=1}^{B_2} y^*(b_2)/B_2$.

c) Estimate the 100α th percentile of the bootstrap-*t* distribution of $W(Y,\theta)$ by the value $\hat{t}_\alpha$ such that $|\{W^*(b_1) \le \hat{t}_\alpha\}|/B_1 = \alpha$.

d) The 95 % “nested bootstrap-t " confidence interval is

$$(y - \hat{t}_{0,975} \cdot u(y) \;\; y + \hat{t}_{0,025} \cdot u(y)).$$

Although it is a more general approach, the nested bootstrap is computationally more involved and harder to implement. The simpler bootstrap method was chosen to analyse all examples.

8.3 Example 1

8.3.1 General

8.3.1.1 As an illustration, consider the statistical model given for Example 1 in Clause 7, namely,

$$Y_i = \theta + \beta + \varepsilon_i, \quad i = 1, ..., n, \tag{14}$$

where θ is the measurand, β represents the background and ε_i are independent $N(0, \sigma^2)$ errors. For a fixed value β, with γ denoting the mean of the data, the measurement equation for this model is $\theta = f(\beta, \gamma) = \gamma - \beta$.

8.3.1.2 If the background, β, has a uniform distribution on the interval $(a - d \;\; a + d)$, the interval for θ derived using the GUM is

$$\overline{Y} - a \pm 2\sqrt{\frac{\sigma^2}{n} + \frac{d^2}{3}}.$$

Reference [5] discusses the properties of such intervals and compares them to the interval

$$\overline{Y} - a \pm \left[2\sqrt{\frac{\sigma^2}{n}} + d \right], \tag{15}$$

which is recommended by Eisenhart [17] and which can be motivated as follows. Since the conditional distribution of $\overline{Y}$ for a given β is Gaussian, $N(\theta + \beta, \sigma^2/n)$,

$$P\left(|\overline{Y} - \theta - \beta| \leq \frac{2\sigma}{\sqrt{n}} \right) \geq 0,95,$$

while

$$P\left(|a - \beta| \leq d \right) = 1.$$

It follows that the Eisenhart interval in (15), namely,

$$P\left(|\overline{Y} - a - \theta| \leq \frac{2\sigma}{\sqrt{n}} + d \right) \geq 0,95, \tag{16}$$

is conservative.

8.3.1.3 However, if $d > 12\sigma/\sqrt{n}$, the interval recommended in the GUM contains interval (15), which demonstrates the difference between these two approaches.

8.3.1.4 The interval (15) can be adjusted for a *t*-distributed ratio $\sqrt{n}(\overline{Y} - a - \beta)/S$. It can also be adjusted for other distributions for the background (triangular, trapezoidal, etc.). Different frequentist methods for

construction of confidence intervals are available in this situation. Indeed in the model (14), $\bar{Y}$ subsumes all the information the data provide about θ (that is, $\bar{Y}$ is a sufficient statistic for θ) with probability density

$$\frac{\sqrt{n}}{2\sqrt{2\pi}\sigma d}\int_{a-d}^{a+d} e^{-0.5n(\bar{y}-\theta-\beta)^2/\sigma^2}\,\mathrm{d}\beta.$$

The special form of this distribution allows alternative confidence intervals (all centred at the maximum likelihood estimator, $\bar{Y}-a$, but of different lengths) to be derived [14].

8.3.2 Example 1a

The simple example introduced in Clause 7 summarizes the measured values in the model (14) by $\bar{y}=3{,}537$ and $u(\bar{y})=0{,}153$. The latter is substituted for $\sigma/\sqrt{n}$ in Inequality (16) and the factor 2 should be replaced by the percentile of the *t*-distribution with 5,15 effective degrees of freedom. In Example 1a, the background β can be estimated from measured values regarded as drawn from a Gaussian distribution, leading to $\bar{b}=1{,}228$ and $u(\bar{b})=0{,}059$. The resulting estimate of θ is $\bar{y}-\bar{b}=2{,}309$ with associated standard uncertainty $\sqrt{u^2(\bar{y})+u^2(\bar{b})}=0{,}164$. The GUM confidence interval is

$$2{,}309\pm 2{,}548\times 0{,}164 = 2{,}309\pm 0{,}417 = (1{,}892\ \ 2{,}727).$$

The 100($1-\alpha$) % bootstrap-*t* confidence interval according to (13) is $(2{,}309-0{,}164\cdot\hat{t}_{1-\alpha/2}\ \ 2{,}309+0{,}164\cdot\hat{t}_{\alpha/2})$, where $\hat{t}_{\beta}$ is the 100β th percentile of W^* of (12).

For the benefit of users of the R-language and WinBUGS, some R-code fragments [18] and WinBUGS fragments [19] are used to illustrate some of the concepts in this Technical Report. For Example 1a, an `R` program for generating the B = 10 000 realizations of W^* is listed below.

```
B = 10000
y.star = rnorm(B, mean=3.537, sd=0.153)
u.y.star = 0.153 * sqrt(rchisq(B, df=4)/4)
b.star = rnorm(B, mean=1.228, sd=0.059)
u.b.star = 0.059 * sqrt(rchisq(B, df=4)/4)
w.star = ((y.star-b.star)-2.309)/sqrt(u.y.star^2+u.b.star^2)
```

The 95 % bootstrap-*t* confidence interval based on the 0,025 and 0,975 quantiles of the simulated distribution is

```
 2.309 - quantile(w.star, c(0.975,0.025))*0.164
## 1.895754 2.728817
```

[2)]

Namely, the 95 % bootstrap-*t* confidence interval is given by (1,896 2,729)[3)]

2) In the examples computed using `R`, WinBugs, or other software packages, the output is given as reported using the software's standard format. As indicated explicitly by the values of the uncertainties reported, not all digits in the output may be significant digits. Note also that the standard output from these software packages uses the period rather than the comma as a decimal indicator.

3) Values rounded to the equivalent of three significant digits in the expanded uncertainty. Note: for Monte Carlo methods, recomputation of examples will be subject to random error from simulation.

8.3.3 Example 1b

When there is no statistical data for the background, β is instead assumed to have a uniform distribution on the interval $(1{,}126 \;\; 1{,}329)$. Then, the approximate confidence interval derived from the use of the GUM is

$$3{,}537 - 1{,}228 \pm 2{,}533\sqrt{\frac{0{,}342^2}{5} + \frac{0{,}102^2}{3}} = 2{,}310 \pm 0{,}415 = (1{,}895 \;\; 2{,}724).$$

The Eisenhart confidence interval is wider, namely,

$$3{,}537 - 1{,}228 \pm \left[2{,}776\frac{0{,}342}{\sqrt{5}} + 0{,}102 \right] = 2{,}310 \pm 0{,}526 = (1{,}783 \;\; 2{,}836).$$

Similar to Example 1a, a bootstrap-*t* confidence interval can be constructed for θ. For this example, the estimates and the associated standard uncertainties for y, β and θ are numerically the same as those in Example 1a, except that β is determined based on experience or expert opinions and its associated uncertainty is obtained by a Type B evaluation. Therefore, the realizations of W^* are generated in a different way from those in from Example 1a, namely, only in generating the bootstrap sample b^* and its associated uncertainty. The bootstrap sample b^* is now generated from the known uniform (1,126, 1,329) distribution with standard uncertainty 0,059. The R code for generating $B = 10\,000$ realizations of W^* is as follows.

```
B = 10000
y.star = rnorm(B, mean=3.537, sd=0.153)
u.y.star = 0.153 * sqrt(rchisq(B, df=4)/4)
b.star = runif(B, min=1.126, max=1.329)
u.b.star = 0.059
w.star = ((y.star-b.star)-2.309)/sqrt(u.y.star^2+u.b.star^2)
```

The 95 % bootstrap-*t* confidence interval based on the 0,025 and 0,975 quantiles of the distribution so formed is

```
2.309 - quantile(w.star, c(0.975,0.025))*0.164
## 1.918643 2.699749
```

Namely, the 95 % bootstrap-t confidence interval is given by (1,919 2,700).

8.3.4 Example 1c

Since $\overline{y} = 1{,}196, s_{\overline{y}} = 0{,}047$, both intervals have negative lower end-points. If the mean θ is known to be positive, these end-points are replaced by zero leading to the GUM-recommended interval $(0 \;\; 0{,}124)$ and to the Eisenhart interval $(0 \;\; 0{,}202)$.

An R program for generating the B = 10 000 realizations of W^* to obtain the bootstrap interval is the same as Example 1b with $\overline{y} = 1{,}196$ and $u(\overline{y}) = 0{,}047$.

```
B = 10000
y.star = rnorm(B, mean=1.196, sd=0.047)
u.y.star = 0.047 * sqrt(rchisq(B, df=4)/4)
b.star = runif(B, min=1.126, max=1.329)
```

```
u.b.star = 0.059
w.star=((y.star-b.star)+0.032)/sqrt(u.y.star^2+u.b.star^2)
```

The untruncated 95 % bootstrap-*t* confidence interval is

```
-0.032-quantile(w.star,c(0.975,0.025))*0.075
## -0.1762648 0.1128422
```

Namely, the 95 % bootstrap-*t* confidence interval is given by (–0,176 0,113).

As θ is known to be positive, the truncated 95 % bootstrap-*t* confidence interval for θ is $(0 \;\; 0{,}113)$.

9 Bayesian approach for uncertainty evaluation

9.1 Basic method

9.1.1 In metrology, the measurand and the input variables of model (1) are physical quantities with fixed quantity values. Nevertheless, under the Bayesian approach, the corresponding parameters μ_i and θ are considered as random variables in the sense that their probability distributions summarize knowledge about these quantities.

9.1.2 The Bayesian framework uses a definition of probability that allows probability distributions to be defined without physical data, for example, using manufacturers' specifications or other expert knowledge. In typical metrology applications, however, there are measured values (data) of physical quantities that can be used to estimate one or more input quantities. In such cases, a probability density function can be obtained for the quantity using Bayes' theorem as follows. Let $p(\mu_i)$ be a probability density function for μ_i as given before physical data is obtained. This function is called the prior density for μ_i. Let Y denote a random variable for which a realization y (data) exists. The probability density $p(y \mid \mu_i)$ for Y is termed a statistical model. Under the Bayesian framework, since μ_i is a random variable, the notation $\mid$ represents the fact that the probability density of Y is conditional (or depends) on μ_i. For a particular realization $p(y \mid \mu_i)$ of Y, viewed as a function of μ_i is called the likelihood function. Applying Bayes' theorem,

$$p(\mu_i \mid y) = \frac{p(y \mid \mu_i)\, p(\mu_i)}{\int p(y \mid \mu_i)\, p(\mu_i)\, d\mu_i} \tag{17}$$

is the posterior density of μ_i that summarizes our knowledge about μ_i after the data y_i was observed.

9.1.3 When no prior knowledge of the μ_i exists, then a so-called non-informative prior distribution [20] is used. In cases when prior information does exist, it is represented by an informative probability distribution. This is one of the mechanisms, under the Bayesian approach, for including information that is used to perform a Type B evaluation of uncertainty. The form of the likelihood function is usually selected based on knowledge of the process that generates the data.

9.1.4 The form of the likelihood function and the prior densities determine the shape of the posterior density. It is important to select the likelihood function and the prior densities carefully and to perform sensitivity analysis of the results with respect to plausible changes in these distributions. For the prior distributions, this may mean comparing the results of using several different densities. A test of appropriateness of the likelihood function (the statistical model that describes the measurement data) is a form of model validation [21], which applies equally to Bayesian, frequentist, and fiducial models.

9.1.5 The definition of measurement uncertainty given in the Introduction can be interpreted in the context of Bayesian statistics as referring to the posterior probability distribution for the measurand θ, that is, the standard uncertainty is the standard deviation of the random variable (quantity) characterized by this probability distribution. To obtain this standard deviation, it is necessary first to find the joint probability distribution of the μ_i, and then apply a change-of-variables formula [14] to derive the distribution for θ. Moments of this distribution can be obtained more simply as follows. For a function $h(\theta)$, obtain the expected value $\mathrm{E}(h(\theta)) = \int ... \int h(f(\mu_1,...,\mu_p)) p(\mu_1,...,\mu_p) d\mu_1 ... d\mu_p$. The corresponding variance can be obtained as $Var(\theta) = E(\theta^2) - [E(\theta)]^2$. Often, the necessary integration is carried out using Monte Carlo methods [20].

9.1.6 When the μ_i are independent random variables, their joint probability distribution is the product of the individual distributions. In many situations, however, the μ_i are not independent such as when the probability distribution for Y is a function of μ_1 and μ_2, that is, $p(y \mid \mu_1, \mu_2)$ is the statistical model and $p(\mu_1, \mu_2) \neq p(\mu_1) p(\mu_2)$. Then, the posterior density for (μ_1, μ_2) is obtained as

$$p(\mu_1, \mu_2 \mid y) = \frac{p(y \mid \mu_1, \mu_2) p(\mu_1, \mu_2)}{\int p(y \mid \mu_1, \mu_2) p(\mu_1, \mu_2) d\mu_1 d\mu_2}.$$

9.1.7 A common situation that leads to such dependence is when the statistical model is a function of θ, as well as some of the μ_i. Both of the examples considered here fall into this category, illustrating the point that under the Bayesian approach, whenever measurement data is available, the process of specifying the related probability distributions requires an appropriate definition of a statistical model. Doing so will automatically lead to the likelihood functions needed for the application of Bayes' theorem and to appropriate posterior densities. The process can be summarized as follows.

a) Identify all measurement data relevant to the physical quantities of interest (parameters).

b) Specify a statistical model (also called an observation model) relating the data to the parameters, which could be the μ_i or sometimes the measurand θ.

c) Specify prior distributions for all parameters involved.

d) Apply Bayes' theorem to obtain posterior distributions for the parameters.

e) Compute the posterior mean and posterior standard deviation of the measurand.

f) Perform sensitivity analysis of the results with respect to plausible changes in the prior distributions.

9.1.8 Where appropriate, a Taylor series approximation and a normality assumption may be used to avoid the numerical computations. Specifically, the Taylor series expansion of $f(\mu_1,...,\mu_p)$ about the expected values of the μ_i together with a normality assumption can be used to state that $f(\mu_1,...,\mu_p)$ is approximately distributed as $N(f(E(\mu_1),...,E(\mu_p)),\omega^2)$, where

$$\omega^2 = \sum_{i=1}^{p} c_i^2 \, \mathrm{Var}(\mu_i) + 2 \sum_{i<j} c_i c_j \, \mathrm{Cov}(\mu_i, \mu_j).$$

$\mathrm{Cov}(\mu_i, \mu_j)$ denotes the covariance of μ_i and μ_j, and the c_i are the partial derivatives of θ with respect to the μ_i evaluated at the expected values of the μ_i.

NOTE Similarly appearing Formulae (6) and (7) are used in 8.1.8, but there the expansion is employed to find an estimate of the variance of the estimator of θ, not of θ itself.

9.2 Example 1

9.2.1 General

This process is now illustrated on Example 1 given in Clause 7. The measurand in this example is denoted by θ. The measurement model as described in 8.3.1.1 is

$$\theta = \gamma - \beta. \tag{18}$$

9.2.2 Example 1a

9.2.2.1 There are two relevant sets of data: (i) five measured values y_i, obtained independently, of signal plus background, and (ii) five measured values b_i, obtained independently, of background alone. Each value in data set (i) is regarded as a realization of a random variable Y_i having a Gaussian distribution with mean $\gamma = \theta + \beta$ and standard deviation σ_Y, and similarly for each value in (ii) but for a random variable B_i with mean β and standard deviation σ_B. Thus, the statistical model for Y_i is

$$Y_i \mid \theta, \beta, \sigma_Y^2 \sim N(\theta + \beta, \sigma_Y^2),$$

and since the five measured values are obtained independently,

$$p(y_1, ..., y_5 \mid \theta, \beta, \sigma_Y) = \left(\frac{1}{\sigma_Y \sqrt{2\pi}} \right)^5 \exp\left\{ -\frac{\sum_{i=1}^{5} (y_i - \theta - \beta)^2}{2\sigma_Y^2} \right\}.$$

9.2.2.2 The statistical model for B_i is

$$B_i \mid \beta, \sigma_B^2 \sim N(\beta, \sigma_B^2),$$

that is,

$$p(b_1 ..., b_5 \mid \beta, \sigma_B) = \left(\frac{1}{\sigma_B \sqrt{2\pi}} \right)^5 \exp\left\{ -\frac{\sum_{i=1}^{5} (b_i - \beta)^2}{2\sigma_B^2} \right\}.$$

9.2.2.3 Since the two sets of observations are mutually independent, the statistical model for Y and B is

$$p(y, b \mid \theta, \beta, \sigma_Y, \sigma_B) = p(b_1, ..., b_5 \mid \beta, \sigma_B) p(y_1, ..., y_5 \mid \theta, \beta, \sigma_Y).$$

9.2.2.4 There are four parameters, θ, β, σ_Y and σ_B, which are to be assigned prior distributions. In this example, there is no additional information about these parameters, other than that they are non-negative, and thus the random variables will be taken as independent. It is desirable for the forms of the prior distributions to have minimal effect on the results of the analysis. Such an effect is obtained by the use of so-called reference priors [20]. For the parameters associated with the means, that is θ and β, such a density can be approximated by

$$\theta \sim \text{Uniform}(0, c), \qquad \beta \sim \text{Uniform}(0, c)$$

with a large value for c. For the scale parameters σ_Y and σ_B, the reference prior densities

$$p(\sigma_Y) = 1/\sigma_Y, \qquad p(\sigma_B) = 1/\sigma_B$$

are improper, that is, they do not integrate to unity. Since this aspect can cause difficulties in numerical computation, a proper density such as

$$\sigma_Y \sim \text{Uniform}(0, c),$$

or

$$\sigma_Y \sim \text{Gamma}(c, c),$$

with large values of c is used. The notation $\text{Gamma}(\phi_1, \phi_2)$ represents a gamma distribution with parameters ϕ_1 and ϕ_2, that is, for a random variable X, this probability density is given by

$$p(x \mid \phi_1, \phi_2) = \frac{\phi_2^{\phi_1}}{\Gamma(\phi_1)} x^{\phi_1 - 1} e^{-x\phi_2}.$$

This completes the specification of the prior distributions.

9.2.2.5 Application of Bayes' theorem results in the joint posterior density for θ, β, σ_Y and σ_B as follows:

$$p(\theta, \beta, \sigma_Y, \sigma_B \mid y, b) = \frac{p(y, b \mid \theta, \beta, \sigma_Y, \sigma_B) p(\theta) p(\beta) p(\sigma_Y) p(\sigma_B)}{\int p(y, b \mid \theta, \beta, \sigma_Y, \sigma_B) p(\theta) p(\beta) p(\sigma_Y) p(\sigma_B) d\theta \, d\beta \, d\sigma_Y \, d\sigma_B}.$$

The posterior density of the measurand θ is obtained by integration as

$$p(\theta \mid y, b) = \int p(\theta, \beta, \sigma_Y, \sigma_B \mid y, b) d\beta \, d\sigma_Y \, d\sigma_B.$$

This posterior distribution summarizes all information about θ available after the measured values were obtained. The expectation of this distribution is taken as an estimate of the physical quantity and the standard deviation of this distribution is used as the standard uncertainty associated with this estimate. It is straightforward to obtain a coverage interval for the measurand from this distribution. This coverage interval is an interval of possible values for θ with a fixed probability. In Bayesian statistics this interval is called a credible interval. In many cases, numerical methods are employed to accomplish the necessary integrations when applying Bayes' theorem. One possible method of making draws from the posterior distribution is *Markov Chain Monte Carlo* (MCMC) [22] using the software WinBUGS[19]. The code for this example, with the uniform prior distributions with c = 100, is as follows

```
Example1a{
theta~dunif(0,100)
beta~dunif(0,100)
gamma <- theta+beta
sigma.Y~dunif(0,1)
sigma.B~dunif(0,1)
tau.Y <- 1/(sigma.Y*sigma.Y)
tau.B <- 1/(sigma.B*sigma.B)
for(i in 1:n){
y[i]~dnorm(gamma,tau.Y)
b[i]~dnorm(beta,tau.B)}
}
```

With the data given in 7.2, for which n = 5, the program produced a posterior mean of θ of 2,309, and a posterior standard deviation of 0,247. A 95 % credible interval for θ is (1,805 2,815). A Bayesian sensitivity analysis with respect to changes in the form of the four prior distributions can be carried out by varying the value of c (see 9.2.2.4), and by substituting the lines

```
tau.Y~dgamma(1,0E-5,1,0E-5)
tau.B~dgamma(1,0E-5,1,0E-5)
```

for the four lines

```
sigma.Y~dunif(0,1)
sigma.B~dunif(0,1)
tau.Y <- 1/(sigma.Y*sigma.Y)
tau.B <- 1/(sigma.B*sigma.B)
```

and comparing the resulting values of posterior mean and standard deviation. The results in this example are robust to such changes.

9.2.3 Example 1b

9.2.3.1 The information about the background parameter β is provided in the form of a probability distribution obtained by a Type B evaluation of uncertainty. In this case, the observation model is only in terms of data set (i) in Example 1a (9.2.2), that is,

$$Y_i \mid \theta, \beta, \sigma_Y^2 \sim N(\theta + \beta, \sigma_Y^2).$$

9.2.3.2 There are now three parameters that are to be assigned prior distributions. For the background parameter β, the prior density is based on the information given in the Introduction, that is,

$$\beta \sim \text{Uniform}(1{,}126, 1{,}329).$$

For θ and σ_Y,

$$\theta \sim \text{Uniform}(0, c), \quad \sigma_Y \sim \text{Uniform}(0, c),$$

with a large value for c.

This completes the specifications for the prior distributions.

9.2.3.3 The WinBUGS code for this example is as follows.

```
Example1b{
theta~dunif(0,100)
beta~dunif(1.126,1.329)
sigma.Y~dunif(0,1)
gamma <- theta+beta
tau.Y <- 1/(sigma.Y*sigma.Y) for(i in 1:n){
y[i]~dnorm(gamma,tau.Y)
}
```

With the data given in the Introduction, this code produces a posterior mean for θ of 2,309, and a posterior standard deviation for θ of 0,232. A 95 % credible interval for θ is (1,832 2,788). A sensitivity analysis of the results is again satisfactory.

9.2.4 Example 1c

9.2.4.1 The only difference from Example 1b is in the actual measured values (which are now close to the background) and so the same model and WinBUGS code can be used. The posterior mean for θ is now 0,069, the posterior standard deviation for θ is 0,067 and the 95 % credible interval for θ is (0,000 0,188). These results are robust to changes in the value of c with the uniform priors. Changing the form of the prior density for σ_Y from uniform to Gamma results in a posterior mean of 0,058, posterior standard deviation of 0,052 and 95 % credible interval of (0,000 0,150). This is a larger change than in the previous examples, and indicates that here, because of the closeness of the data to the background, the data is not quite as informative about the measurand. The size of σ_Y (controlled to some degree by the prior distribution since there are only five measured values on which an estimate is based) affects how informative is the data. In a case such as this, the conservative solution is to use the longer credible interval based on the uniform distribution. A better way would be to obtain more measured values. The consequence would be a reduction in the effect of the prior density of σ_Y on the results. (An interesting fact about the Bayes' credible intervals such as those given here can be found in Reference [23]. The authors show that in models such as Example 1, the 95 % Bayes' credible interval based on the uniform prior has frequentist coverage of close to 95 %, while the interval based on the gamma prior usually has lower frequentist coverage.)

9.2.5 Summary of example

Example 1a illustrates the case when measured values from two independent sources are used in a single uncertainty evaluation. Example 1b shows how information about the background that is used to perform a Type B evaluation of uncertainty can be included in the Bayesian model. Example 1c illustrates the ease with which a constraint can be included in the Bayesian model, such as the positive constraint here on the value of the measurand. It also shows how the choice of a non-informative prior distribution can affect the results.

10 Fiducial inference for uncertainty evaluation

10.1 Basic method

10.1.1 For the measurement function (1), the uncertainty evaluation for a measurand θ may be based on the fiducial distribution for θ. The following examples serve to illustrate the recipe for obtaining fiducial distributions for parameters of interest.

10.1.2 Suppose $Y \sim N(\theta,1)$, where θ is the measurand, the measurement process has a known variance equal to 1, and Y is the random variable representing values that may be observed. One might express the relation between the measured values and the underlying random experimental error process by the equation

$$Y = \theta + E, \tag{19}$$

where *E* is a random error with $N(0,1)$ distribution. Each measured value is associated with a particular random experimental error. Suppose a single measured value of 10 is obtained. The associated measurement error is denoted by e. So

$$10 = \theta + e.$$

Hence $\theta = 10 - e$. If the value of e were known, the measurand would be known exactly, but the value of e is not known. Nevertheless, the fact that the distribution from which *e* was generated is known helps a set of values of θ to be determined that is considered plausible. For instance, how plausible is the value $\theta = 2$ for the measurand? For this to be true, e = 8 is needed. A value of 8 is highly unlikely to have come from an $N(0,1)$ distribution. So, it is concluded that the value $\theta = 2$ is unlikely. How likely is it that θ lies between

10 and 12? For θ to be between 10 and 12, e needs to be between 0 and 2 and the probability for this is $\Phi(2)-\Phi(0)$, where $\Phi(z)$ is the value of the cumulative standard Gaussian distribution at z. Thus, probabilities associated with E can be transferred to probabilities for θ. The knowledge about θ, based on the measured value of 10, can be described by the distribution of the random variable $\tilde{\theta}$ whose distribution is given by that of 10 – E. That is, $\hat{\theta} \sim N(10,1)$ or the fiducial distribution for θ (that is, the distribution for $\tilde{\theta}$) is $N(0,1)$. The random variable $\tilde{\theta}$ is also called a fiducial quantity (FQ) for θ. Such an FQ is related to what is called generalized pivotal quantity [24], [25] or fiducial generalized pivotal quantity [26], [27] in the literature.

10.1.3 In the above example, suppose two measurements are made. Let Y_1 and Y_2 be the random variables denoting the possible values one might obtain for the two measurements, which can be expressed as

$$Y_1 = \theta + E_1,$$

$$Y_2 = \theta + E_2. \tag{20}$$

Suppose the actual measured values are 10 and 8. Then, the following equations relate the measured values, the measurand, and the realized values of experimental errors, say e_1 and e_2

$$10 = \theta + e_1,$$
$$8 = \theta + e_2.$$

Plausible values for θ are related to plausible values of (e_1, e_2). What makes this example different from the previous example is that here it is known that $e_1 - e_2$ equals 2. So, the set of possible values for (e_1, e_2) is now limited by this requirement. It is known that (e_1, e_2) is from a standard bivariate Gaussian distribution, but is constrained to lie on the line $e_1 - e_2 = 2$. So, the probabilities one would associate with θ are the probabilities with either 10 – e_1 or 8 – e_2 knowing that (e_1, e_2) is a realization from a bivariate standard Gaussian distribution subject to the additional condition that $e_1 - e_2 = 2$. Hence an FQ $\tilde{\theta}$ is defined to have a distribution that is equal to the conditional distribution of $10 - E_1$ given that $E_1 - E_2 = 2$. This is the same distribution as the conditional distribution of $8 - E_2$ given that $E_1 - E_2 = 2$. A simple calculation shows that the distribution of $\tilde{\theta}$ is $N(\bar{y}, 1/2)$ where $\bar{y} = (y_1 + y_2)/2 = (10+8)/2 = 9$.

10.1.4 More generally, for n independent measurements made from $N(\theta, \sigma^2)$,

$$Y_1 = \theta + \sigma E_1,$$

$$Y_2 = \theta + \sigma E_2,$$

…,

$$Y_n = \theta + \sigma E_n, \tag{21}$$

where $E_1, \ldots, E_n$ are independent, standard Gaussian random variables. The joint fiducial distribution for (θ, σ) can be obtained as follows. Use the first two (or any two) of the above n structural equations to solve for θ and σ, denoted by $\tilde{\theta}$ and $\tilde{\sigma}$, as functions of y_1, y_2, E_1, and E_2. The joint fiducial distribution for (θ, σ) is the joint distribution for $(\tilde{\theta}, \tilde{\sigma})$ conditioned on the E_i imposed by the rest of the $n-2$ equations. In particular, the fiducial distribution for θ is

$$\tilde{\theta} = \bar{y} - \frac{s}{\sqrt{n}} T_{n-1}, \tag{22}$$

namely, a shifted and scaled *t*-distribution with $n-1$ degrees of freedom. Here $\bar{y}$ and s are the realized values of the sample mean $\bar{X}$ and the sample standard deviation S for the n measured values, and T_{n-1} is a random variable having a *t*-distribution with $n-1$ degrees of freedom.

10.1.5 There is an alternative, simpler method than that just outlined to derive a fiducial distribution for θ in (22), which will be illustrated in the subsequent examples.

10.1.6 The above argument can be generalized and fiducial distributions can be developed for model parameters in wide-ranging problems. The starting point for this process is called a *structural equation* [28]. Denote this structural equation by $Y = G(\beta, E)$. For a single measurement, Equation (19) constitutes the structural equation. For n measurements, equations (21) constitute the structural equations. The structural equations relate the measurements Y with model parameters β and error processes E whose distributions are fully known. For instance, for a single measurement the distribution for E is known completely. For any fixed values of β, the distribution for E and the structural equation $G(\cdot)$ determine the distribution for the data Y. After observing the data Y the role of data and parameters can be interchanged. In particular, the value of Y is fixed and the distribution of E and the structural equation $G(\cdot)$ are used to infer a distribution for β. This is what constitutes the fiducial argument.

10.2 Example 1

10.2.1 Example 1a

10.2.1.1 To illustrate, consider Example 1a described in Clause 7 where the physical quantity θ is to be estimated from measured values that follow the model

$$Y_i = \theta + \beta + \varepsilon_i, \quad i = 1, ..., n, \tag{23}$$

where the ε_i are independent measurement errors with $\varepsilon_i \sim N(0, \sigma_Y^2)$. Also, β represents a background and can be estimated from measurements that follow the model

$$B_i = \beta + \delta_i, \quad i = 1, ..., n_b, \tag{24}$$

where δ_i are independent measurement errors with $\delta_i \sim N(0, \sigma_B^2)$. It is assumed that ε_i and δ_i are independent. From (23) and (24), it follows that $\bar{Y} - \bar{B}$ has a Gaussian distribution with mean θ and variance $\sigma_Y^2/n + \sigma_B^2/n_b$, where $\bar{Y}$ and $\bar{B}$ are the means of Y_i and B_i, respectively and can be expressed by

$$\bar{Y} - \bar{B} = \theta + \sqrt{\frac{\sigma_Y^2}{n} + \frac{\sigma_B^2}{n_b}}\, Z, \tag{25}$$

where Z is a standard Gaussian random variable. This is a structural equation for $\bar{Y} - \bar{B}$. Also

$$W_y = \frac{(n-1)S_y^2}{\sigma_Y^2} \sim \chi^2(n-1)$$

and

$$W_b = \frac{(n_b - 1)S_b^2}{\sigma_B^2} \sim \chi^2(n_b - 1),$$

where $\chi^2(\nu)$ stands for the chi-squared distribution with ν degrees of freedom, S_y^2 and S_b^2 are sample variances of Y_i and B_i, respectively. Thus

$$S_y^2 = \frac{\sigma_Y^2 W_y}{n-1} \tag{26}$$

is a structural equation for S_y^2 and

$$S_b^2 = \frac{\sigma_B^2 W_b}{n_b - 1} \tag{27}$$

is a structural equation for S_b^2. By solving the above three structural equations for θ, σ_Y, and σ_B, an FQ for θ is obtained as

$$\tilde{\theta} = \overline{y} - \overline{b} - \sqrt{\frac{(n-1)s_y^2}{nW_y} + \frac{(n_b-1)s_b^2}{n_b W_b}}\, Z. \tag{28}$$

10.2.1.2 A $1-\alpha$ fiducial interval for θ is given by $(\tilde{\theta}_{\alpha/2} \;\; \tilde{\theta}_{1-\alpha/2})$, where $\tilde{\theta}_\alpha$ is the α-quantile of the distribution of $\tilde{\theta}$. These quantiles can be determined analytically in simple situations. However, they are most conveniently approximated using a Monte Carlo approach. This approach involves generating a large number of realizations from the distribution of $\tilde{\theta}$ and determining the empirical $\alpha/2$ and $1-\alpha/2$ quantiles. These quantiles are used as the estimates for $\tilde{\theta}_{\alpha/2}$ and $\tilde{\theta}_{1-\alpha/2}$. A single realization of $\tilde{\theta}$ may be generated as follows.

a) Generate a realization of a standard Gaussian random variable Z.

b) Generate realizations of independent χ^2 random variables W_y and W_b with $n-1$ and n_b-1 degrees of freedom, respectively.

c) Calculate $\tilde{\theta}$ as in (28).

For this example, $n = n_b = 5$, $\overline{y} = 3{,}537$, $s_y = 0{,}342$, $\overline{b} = 1{,}228$, and $s_b = 0{,}131$. An R program for generating the 500 000 realizations of $\tilde{\theta}$ is listed below.

```
nrun = 500000
Z = rnorm(nrun)
W1 = rchisq(nrun, 4)
Wb = rchisq(nrun, 4)
theta = 3.537 - 1.228 - sqrt(4*0.342^2/(5*W1)+4*0.131^2/(5*Wb))*Z
```

The mean of the simulated distribution is

```
mean(theta)
## 2.308893
```

and a 95 % fiducial interval based on the 0,025 and 0,975 quantiles of the simulated distribution is

```
quantile(theta, c(0.025,0.975))
##  2.5 %  97.5 %
## 1.857814  2.760931
```

Namely, the 95 % fiducial interval is given by (1,858 2,761).

10.2.2 Example 1b

10.2.2.1 There is now no statistical data relating to the background. It is assumed that the information regarding β is specified in terms of a probability distribution for β and that β and ε_i are independent. Furthermore, it is assumed the probability distribution for β is fully known, that is, does not involve any unknown parameters.

10.2.2.2 The structural equation for $\overline{Y}$ is given by

$$\overline{Y} = \theta + \beta + \frac{\sigma}{\sqrt{n}} Z. \tag{29}$$

Together with the structural equation for S_y^2 in (26), we obtain an FQ for θ as

$$\tilde{\theta} = \overline{y} - \beta - \frac{s_y}{\sqrt{n}} \frac{Z}{\sqrt{W_y/(n-1)}}.$$

Since $Z/\sqrt{W_y/(n-1)} = T_{n-1}$ is a random variable having a *t*-distribution with $n-1$ degrees of freedom,

$$\tilde{\theta} = \overline{y} - \beta - \frac{s_y}{\sqrt{n}} T_{n-1}. \tag{30}$$

A single realization of $\tilde{\theta}$ may be generated as follows.

a) Generate a realization of T_{n-1} of a Student's *t* random variable with $n-1$ degrees of freedom.

b) Generate β according its distribution, independently of T_{n-1}.

c) Calculate $\tilde{\theta}$ as in (30).

For this example, β is assumed uniformly distributed over the interval (1,126 1,329). The 500 000 realizations of $\tilde{\theta}$ are generated by

```
beta = runif(nrun, 1.126, 1.329)
theta = 3.537 - beta - 0.342/sqrt(5)*rt(nrun, 4)
```

The mean of the simulated distribution is

```
mean(theta)
## 2.309454
```

and a 95 % fiducial interval based on the 0,025 and 0,975 quantiles of the simulated distribution is

```
quantile(theta, c(0.025, 0.975))
## 2.5 %    97.5 %
## 1.871685 2.745590
```

Namely, the 95 % fiducial interval is given by (1,872 2,746).

10.2.2.3 The above fiducial interval agrees with the uncertainty interval obtained using the method proposed in GUMS1.

10.2.3 Example 1c

10.2.3.1 The case of Example 1b applies except $\bar{y} = 1{,}196$ and $s_y = 0{,}106$. The 500 000 realizations of $\tilde{\theta}$ are generated by

```
theta = 1.196 - beta - 0.106/sqrt(5)*rt(nrun, 4)
```

The mean of the realizations is

```
mean(theta)
## -0.03158058
```

which lies outside the parameter space for θ. The number of realizations outside of the parameter space can be found by

```
length((1:nrun)[theta < 0])
## 319168
```

The approach for handling parameter constraints is to truncate the fiducial distribution to the constrained parameter space. That is, we use $\max(\tilde{\theta}, 0)$ to obtain the realizations of the fiducial distribution for θ. A 95 % fiducial interval is calculated as

```
quantile(pmax(theta, 0), c(0.025, 0.975))

##  2.5 %    97.5 %
##  0.0000000 0.1361553
```

Namely, the 95 % fiducial interval is given by (0,000 0,136).

10.2.3.2 The recipe described in 10.2.1.1 and 10.2.2.2 can be generalized to arbitrary statistical models. A prescription for constructing FQs is given in Reference [29]. A simpler recipe for more common problems where sufficient statistics exist was given in a technical report (Reference [30]) and is further discussed in References [24] and [25]. It is reproduced here for completeness. The recipe consists of the following steps:

a) Express each sufficient statistic as a function of one or more parameters and random variables whose distributions are completely known, free of any unknown parameters. That is, obtain a structural equation for each sufficient statistic.

b) In each structural equation, express each parameter as a function of the sufficient statistics and random variables whose distributions are completely known.

c) Obtain an FQ for each parameter by replacing the sufficient statistics with their corresponding observed values.

11 Example 2: calibration of a gauge block

11.1 General

11.1.1 This example, which is taken from Annex H.1 of the GUM, is concerned with the determination of the length of a gauge block by comparing it with a nominally identical gauge bock that has previously been calibrated. The notation used in the GUM is closely followed, but has been modified where needed to agree with the notational conventions in Clause 4 that distinguish measurands from measured values. Table 2 lists the physical quantities used.

11.1.2 Use the notation given in Table 3 and based on

— the first line of Equation (H.2) in the GUM,

— the relationships $\alpha = \alpha_s + \delta_\alpha$ and $\theta_s = \overline{\theta} + \Delta - \delta_\theta$ defined in H.1.2, and

— inferences drawn from the propagation of uncertainty in H.1.3.2 and H.1.3.4 of the GUM.

11.1.3 The measurement model for λ used in the GUM analysis of Example H.1 can be expressed as

$$\lambda = \frac{\lambda_s\left[1+\alpha_s\left(\overline{\theta}+\Delta-\delta_\theta\right)\right]+\delta_\lambda+\delta_{C_{\mathrm{r}}}+\delta_{C_{\mathrm{nr}}}}{1+\left(\alpha_s+\delta_\alpha\right)\left(\overline{\theta}+\Delta\right)}. \tag{31}$$

Equation (31) is the measurement model as described in H.1.1 and the first line of Equation (H.2) of the GUM, rather than the approximation made on the second line of Equation (H.2) and then used throughout the rest of H.1.

Table 3 — Notation for analysis of GUM Example H.1 under each of the three statistical approaches. The random variable corresponding to δ_λ is denoted by $\overline{D}_\lambda$, and its observed value by $\overline{d}_\lambda$

Quantity	Symbol
Length of unknown end gauge at 20 °C	λ
Length of standard end gauge at 20 °C	λ_s
Difference between end gauge lengths at laboratory ambient temperature	δ_λ
Correction to difference between gauge block lengths to compensate for random comparator errors	$\delta_{C_{\mathrm{r}}}$
Correction to difference between gauge block lengths to compensate for systematic comparator errors	$\delta_{C_{\mathrm{nr}}}$
Coefficient of thermal expansion of the standard end gauge	α_s
Difference in coefficients of thermal expansion of the standard and unknown end gauges	δ_α
Average deviation of test bed temperature from standard conditions during data collection	$\overline{\theta}$
Cyclic variation of test bed temperature from mean temperature due to thermostatic control	Δ
Difference in temperatures of the standard and unknown end gauges	δ_θ

11.1.4 Equation (31) is also expressed in terms of the physical quantities used to determine the length of the gauge block, rather than pre-summarizing the effects due to the difference between the lengths of the two gauges and in the temperature of the test bed. It is good practice to express the measurement model in terms of all quantities needed to determine it. This practice helps to minimize a possible failure to identify correlations between different physical quantities, such as θ and θ_s and α and α_s as mentioned in H.1.2, whose values ultimately might be based on the same data.

11.1.5 Table 4 summarizes the rest of the information taken from the analysis of GUM example H.1 needed for the analysis by the different statistical approaches to be discussed and compared in the remainder of this clause.

11.1.6 The description of the example in the GUM indicates that there is only one quantity, δ_λ, whose value has been estimated using the analysis of statistical data. The distribution of the mean of the measured values, which provides an estimate of δ_λ, is taken as Gaussian (normal) with an expected value that depends on the length of the gauge block and the other physical quantities described in Table 3 and Table 4.

11.1.7 The values and standard uncertainties associated with the estimates of all other quantities are obtained by Type B evaluations. Because the quantities δ_α and δ_θ follow rectangular rather than Gaussian distributions, however, there are no widely accepted statistical methods to account for the degrees of freedom in these two cases. As a result, the given degrees of freedom will not be used for those quantities.

Table 4 — Summary of information from the analysis of GUM example H.1 needed for its re-analysis

Quantity	Value	Standard uncertainty	Degrees of freedom	Type of uncertainty evaluation	Characterizing distribution
λ_s	50 000 623 nm	25 nm	18	B	Gaussian
$\overline{d}_\lambda$	215 nm	5,8 nm	24	A	Gaussian
δ_{C_r}	0 nm	3,9 nm	5	B	Gaussian
$\delta_{C_{nr}}$	0 nm	6,7 nm	8	B	Gaussian
α_s	$11{,}5\times10^{-6}\ ^\circ\mathrm{C}^{-1}$	$1{,}2\times10^{-6}\ ^\circ\mathrm{C}^{-1}$		B	Rectangular
δ_α	$0\ ^\circ\mathrm{C}^{-1}$	$0{,}58\times10^{-6}\ ^\circ\mathrm{C}^{-1}$	50	B	Rectangular
$\overline{\theta}$	$-0{,}1\ ^\circ\mathrm{C}$	$0{,}2\ ^\circ\mathrm{C}$		B	Not specified
Δ	$0\ ^\circ\mathrm{C}$	$0{,}35\ ^\circ\mathrm{C}$		B	Arcsine
δ_θ	$0\ ^\circ\mathrm{C}$	$0{,}029\ ^\circ\mathrm{C}$	2	B	Rectangular

11.2 Frequentist approach

11.2.1 In this example, the sensitivity coefficients $c_{\alpha_s} = c_{\theta_s}$ vanish, and the second order terms are to be incorporated in Equations (6) and (7), although just one of them is noticeably different from zero (GUM, p 71).

11.2.1 The value y = 50 000 838 nm of the measurand, namely, the length of the gauge block under calibration, and the associated standard uncertainty $u(y) = 34$ nm are returned by the evaluation in the GUM using second-order terms. As was mentioned in 8.1.12, the uncertainties associated with these estimates are approximated by the marginal quadratic error if the parameters λ_s, θ_s are averaged over their normal distributions, and λ_s, θ_s and δ_θ are integrated out according their uniform distributions.

11.2.2 These results are confirmed by the propagation of distributions implemented by a Monte Carlo method as in GUMS1, which provides a very close answer. Moreover, the approximation by the *t*-distribution (10) seems to be reasonable. Figure 1 shows the empirical percentiles plotted against *t*-distribution quantiles when the degrees of freedom are estimated according to Equation (8). More results from a Monte Carlo method are reported in Reference [31].

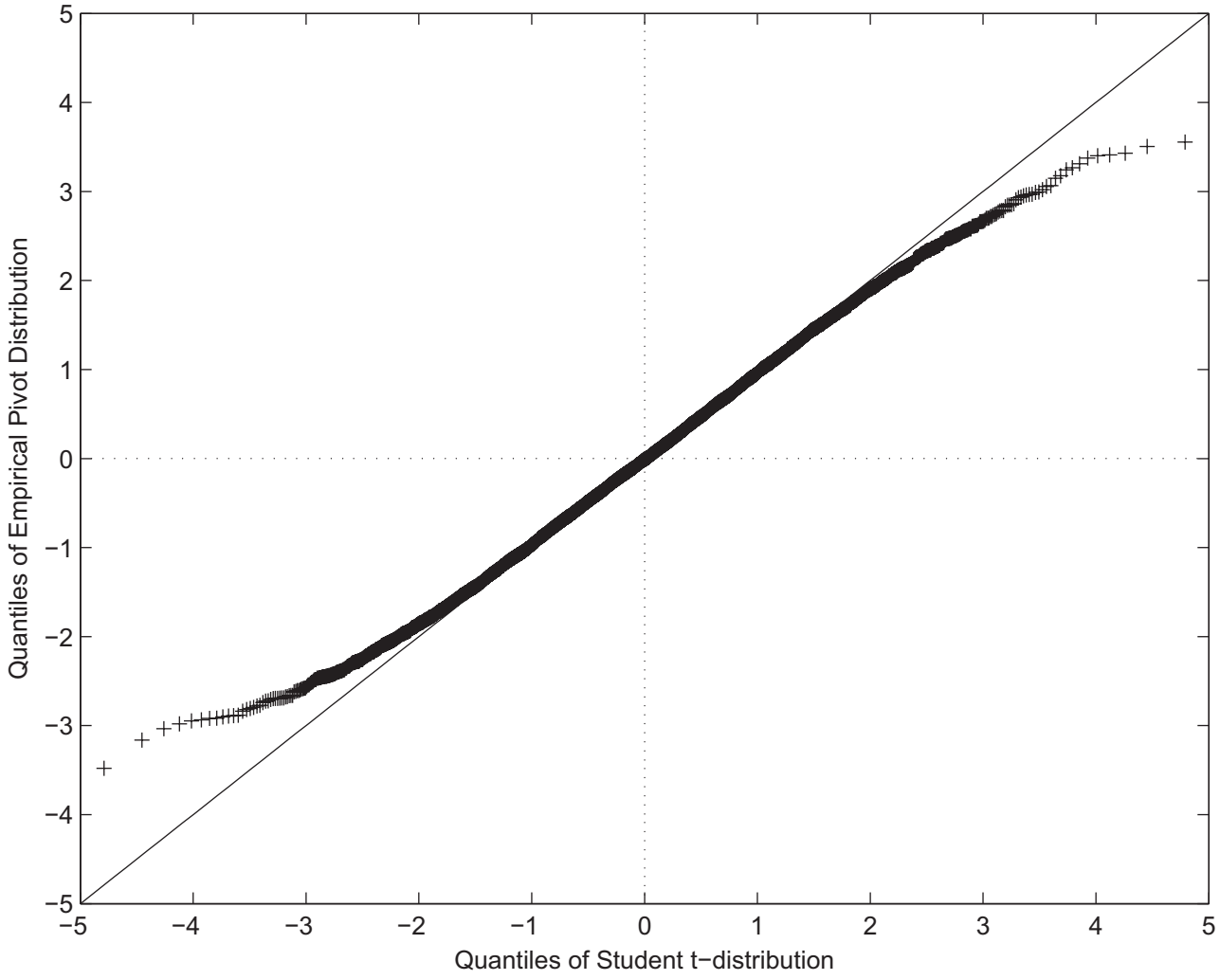

Figure 1 — Empirical percentiles versus percentiles of *t*-distribution in Example 2

11.2.3 To construct a bootstrap interval for this example, λ is estimated to be 50 000 838 nm with a combined standard uncertainty u = 31,7 nm. From Equation (13), the 100(1 – α) % bootstrap-*t* confidence interval is $(50\,000\,838 - \hat{t}_{1-\alpha/2} \cdot 31{,}7 \;\; 500\,00\,838 + \hat{t}_{\alpha/2} \cdot 31{,}7)$, where $\hat{t}_{\beta}$ is the 100β th percentile of W^* of (12). The R code for generating B = 10 000 realizations of W^* is as follows.

```
B = 10000
x.star = cbind(
     rnorm(B, mean=50000623, sd=25),
     rnorm(B, mean=215, sd=5.8),
     rnorm(B, mean=0, sd=3.9),
     rnorm(B, mean=0, sd=6.7),
     runif(B, min=0.0000095, max=0.0000135),
     runif(B, min=-0.000001, max=0.000001),
     runif(B, min=-0.45, max=0.25),
     rbeta(B, 0.5, 0.5)-0.5,
     runif(B, min=-0.05, max=0.05))
```

```
u.star = cbind(
    25 * sqrt(rchisq(B, df=18)/18),
    5.8 * sqrt(rchisq(B, df=24)/24),
    3.9 * sqrt(rchisq(B, df=5)/5),
    6.7 * sqrt(rchisq(B, df=8)/8),
    0.000001 2,
    0.00000058 * sqrt(rchisq(B, df=50)/50),
    0.2,
    0.35,
    0.029 * sqrt(rchisq(B, df=2)/2))
x.name = c("L.s","D.lambda","Dc.r","Dc.s","A.std","D.alpha",
    "T.bar","T.cv","D.theta")
f = expression((L.s*(1+A.std*(T.bar+T.cv-D.theta))+
    D.lambda+Dc.r+Dc.s)/ (1+(A.std+D.alpha)*(T.bar+T.cv)))
star = delta(f, x.star, u.star, x.name)
w.star = (star$y-50000838)/star$uc
```

The R function delta is defined below.

```
delta = function(meq,x,u,namevec){
    for(i in 1:ncol(x)) assign(namevec[i], x[,i])
    c = attr(eval(deriv(meq,namevec)),"gradient")
list(y=eval(meq),uc=sqrt(apply((c*u)^2,1,sum)))
 }
```

This function accepts a valid R expression meq, the measurement function, whose parameter names are given by namevec, and a matrix of input values x, one column of x containing the bootstrap replicates for each quantity in meq. The function uses the R function deriv to evaluate the measurement function and obtain the first derivatives c (the "gradient" of meq at x) with respect to all the parameters in namevec evaluated at the input values given by x. Finally, the function returns the evaluated expression (as y) and the associated uncertainty calculated using the usual first-order Taylor approximation. In the bootstrap code above, the function delta is applied to the measurement function defined as f in the code.

The 95 % bootstrap-*t* confidence interval based on the 0,025 and 0,975 quantiles of the simulated distribution is

```
50000838-quantile(w.star,c(0.975,0.025))*31.70511
## 50000777 50000899
```

This interval, (50 000 777 50 000 899) nm, is almost 10 % shorter than that given by application of the GUM. This general behaviour, the width of a bootstrap interval being shorter than that of an interval derived from an uncertainty evaluation based on the first-order Taylor approximation, is further discussed in Reference [32].

11.3 Bayesian approach

11.3.1 An observational model relating the data to the parameters can be specified.

11.3.2 The GUM can be interpreted as stating that the expected value $E(D_\lambda)$ of the measurement is equal to δ_λ, where

$$\delta_\lambda = \lambda\left(1+(\delta_\alpha+\alpha_s)\left(\overline{\theta}+\Delta\right)\right)-\lambda_s\left(1+\left(\left(\overline{\theta}+\Delta\right)-\delta_\theta\right)\alpha_s\right). \tag{32}$$

The expected value of the measurements is a function of the parameter vector $\gamma=(\lambda_s,\overline{\theta},\Delta,\delta_\alpha,\alpha_s,\delta_\theta)$ and the measurand λ. The GUM gives two additional components of uncertainty involving the comparator determined by Type B evaluations. Therefore there is uncertainty about the expected value of the difference measurement being equal to δ_λ. Similarly to the GUM, the two components can be combined additively to obtain an uncertainty of 7,8 nm, with 12 degrees of freedom using the Welch-Satterthwaite formula. The following two-stage statistical model combines the available information:

$$\overline{D}_\lambda \mid \delta_{\lambda_r} \sim N\left(\delta_{\lambda_r}, \sigma^2_{D_\lambda}/5\right) \tag{33}$$

$$\delta_{\lambda_r} \mid \delta_\lambda \sim \delta_\lambda + 7.8\cdot T_{12}\,.$$

11.3.3 Given in the example is the uncertainty associated with the measured difference s_{d_λ} obtained by a Type A evaluation, providing an estimate of σ_{D_λ}. From basic probability theory, for a sample of size n from a Gaussian distribution with a known variance σ^2,

$$\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2(n-1).$$

As $\chi^2(n-1)$ is also the $\text{Gamma}\left(\frac{n-1}{2},\frac{1}{2}\right)$ density,

$$S^2_{d_\lambda} \mid \sigma_{D_\lambda} \sim \text{Gamma}\left(\frac{(25-1)}{2},\frac{1}{2\sigma^2_{D_\lambda}}\right).$$

11.3.4 There are eight parameters in the statistical model, including the measurand λ. To find a posterior distribution for λ, the joint prior distribution of the eight parameters is first specified. A priori, these random variables can be regarded as independent, and so their joint distribution is the product of their individual prior distributions. For the elements of the parameter vector γ, the information that is used to perform a Type B evaluation of uncertainty can be interpreted as informative prior densities as follows

$$\lambda_S \sim N(50000623,625)\,, \tag{34}$$

$$\delta_\alpha \sim \text{Uniform}\left(-1\times10^{-6},1\times10^{-6}\right),$$

$$\overline{\theta} \sim N(-0{,}1,\ 0{,}1681)\,,$$

$$\Delta \sim \text{Beta}(0{,}5,\ 0{,}5)-0{,}5\,,$$

$$\alpha_S \sim \text{Uniform}\left(9{,}5\times10^{-6},\ 13{,}5\times10^{-6}\right),$$

$$\delta_\theta \sim \text{Uniform}(-0{,}05,\ 0{,}05).$$

So, the joint prior distribution is

$$p(\gamma)=p(\lambda_s)p(\delta_\alpha)p(\overline{\theta})p(\Delta)p(\alpha_s)p(\delta_\theta).$$

Prior distributions for the measurand λ and σ_{D_λ} are needed to complete the prior specification. In this example, there is no additional information about these two parameters other than that they are non-negative. As in Example 1, the parameters are assigned *reference* priors [20]. For λ, a reference density is approximated as

$$\lambda \sim \text{Uniform}(0, c) \tag{35}$$

with a large value for c. Similarly, for σ_{D_λ},

$$\sigma_{D_\lambda} \sim \text{Uniform}(0, c), \tag{36}$$

or $\text{Gamma}(c, c)$. This completes the prior distribution specification.

Note that the two reference prior distributions, which sensitivity analysis shows have little impact on the results, are the only distributions not used in some manner by the frequentist or fiducial approaches.

11.3.5 Application of Bayes' theorem results in the following joint posterior density of $\left(\lambda, \gamma, \sigma_{D_\lambda}\right)$:

$$p\left(\lambda, \gamma, \sigma_{D_\lambda} \mid \overline{d}_\lambda, s_{d_\lambda}\right) = \frac{p(\overline{d}_\lambda \mid \delta_{\lambda_r}) p(\delta_{\lambda_r} \mid \delta_\lambda) p(\gamma) p(\lambda) p(\sigma_{D_\lambda})}{\int p(\overline{d}_\lambda \mid \delta_{\lambda_r}) p(\delta_{\lambda_r} \mid \delta_\lambda) p(\gamma) p(\lambda) d\gamma d\lambda d\sigma_{D_\lambda}}.$$

The posterior density of λ is then obtained by integration as

$$p(\lambda \mid \overline{d}_\lambda, s_{d_\lambda}) = \int p(\lambda, \gamma, \sigma_{D_\lambda} \mid \overline{d}_\lambda, s_{d_\lambda}) d\gamma d\sigma_{D_\lambda}.$$

This posterior distribution summarizes all of the information about λ available after the measurements were obtained. The WinBUGS code for this example is as follows:

```
Example2 {
   n<-25
   df<-(n-1)/2
   lambda~dnorm(0.1,0E-18)
   delta.a~dunif(-0.000001, 0.000001)
   alpha~dunif(0.0000095,0.0000135)
   theta~dnorm(-0.1,5.94)
   ddelt~dbeta(0.5,0.5)
   delta<-ddelt-0.5
   delta.t~dunif(-0.05,0.05)
   lambda.s~dnorm(50000623, 0.0016)
   sigma.D~dunif(0.20)
   tau.D<-1/(sigma.D*sigma.D)
   delta.l<-lambda*(1+(delta.a+alpha)*(theta+delta))
        -lambda.s*(1+((theta+delta)-delta.t)*alpha)
        delta.l.r~dt(delta.l, 0.0164,12)
   msg<-5*tau.D
   dbar~dnorm(delta.l.r,msg)
```

```
pg<-tau.D/2
ssq<-(n-1)*s.y*s.y
ssq~ dgamma(df,pg)
}
```

For $\bar{d}_\lambda$ = 215 and s_Y = 13, this WinBUGS code obtains the posterior mean of λ as 50 000 837 nm, with posterior standard deviation of 34 nm. The 95 % credible interval is (50 000 768 nm 50 000 908 nm). These results are almost identical those in the GUM.

11.3.6 In the solution here, the measurement model in terms of λ, that is Equation (31), is never used, so avoiding the unnecessary and difficult task of determining how the distributions of the various parameters are related. As in Example 1 with the two parameters, the approach given here leads to an appropriate joint posterior distribution for all eight parameters.

11.3.7 Consider an approximate solution for this example based on the Taylor series approximation. In the GUM solution, Equation (31) is approximated as

$$\delta_\lambda = \lambda - \lambda_s\left(1-\left(\delta_\alpha\left(\bar{\theta}+\Delta\right)+\alpha_s\delta_\theta\right)\right).$$

Define a parameter $\eta = \lambda - \delta_\lambda$. Using the Taylor series approximation, the probability density of η can be approximated by a Gaussian as $\eta \sim N$ (50 000 623 nm, 911 nm^2). For simplicity, also approximate σ_{D_λ} by s_{d_λ}. Then, the statistical model becomes

$$\bar{D}_\lambda \mid \delta_{\lambda_r} \sim N\left(\delta_{\lambda_r}, \frac{(13)^2}{5}\right),$$

$$\delta_{\lambda_r} \mid \delta_\lambda \sim N\left(\lambda-\eta, (7{,}8)^2\right),$$

$$\lambda \sim N(0,c),$$

$$\eta \sim N\left(50\,000\,623,\ 911{,}47\right).$$

For this model, the posterior density of λ can be obtained analytically [33]. We obtain

$$\lambda \sim N\left(\bar{d}+50\,000\,623, \frac{(13)^2}{5}+(7{,}8)^2+911{,}47\right).$$

Since $\bar{d}_\lambda$ =215 nm, we obtain the posterior mean of λ as 50 000 838 nm with posterior standard deviation of 31 nm, again close to those in the GUM.

11.4 Fiducial approach

11.4.1 This example is used to illustrate the fiducial inference approach in a more complex application. The measurement function is given in (31). Based on the information provided in the GUM, the following assumptions are made:

a) The estimated value of λ_s (i.e., the value given in the calibration certificate), denoted by l_s, is equal to 50 000 623 nm. The standard uncertainty associated with the estimate is 25 nm with 18 degrees of freedom. Under a normality assumption, a fiducial quantity (FQ) for λ_s is

$$\tilde{\lambda}_s = 50000623 - 25T_{18}. \quad (37)$$

Expression (37) is obtained from Distribution (22) with $\bar{y} =$ 50 000 623 nm, $u(\bar{y})$ = 25 nm, and 18 degrees of freedom associated with $u(\bar{y})$.

b) Each replicate measured value is regarded as drawn from a normal distribution with mean δ_λ and standard deviation σ_{δ_λ}. The observed mean of the five measured values is $\bar{d}_\lambda$ = 215 nm. From a separate experiment, σ_{δ_λ} is estimated to be 13 nm with 24 degrees of freedom. Thus, $u(\bar{d}_\lambda) = 13/\sqrt{5}$. So, an FQ for δ_λ is

$$\tilde{\delta}_\lambda = 215 - 13T_{24}/\sqrt{5}. \quad (38)$$

Also, based on the calibration certificate for the comparator device, an estimate of δ_{C_r} is 0 with a standard uncertainty of 3,9 nm (5 degrees of freedom), and an estimate of $\delta_{C_{nr}}$ is 0 with a standard uncertainty of 6,7 nm (8 degrees of freedom). Furthermore, the comparator errors can be assumed to be independent of the replication errors. Thus,

$$\tilde{\delta}_{C_r} = 3,9T_5, \quad (39)$$

and

$$\tilde{\delta}_{C_{nr}} = 6,7T_8. \quad (40)$$

Mutual independence among the random variables is a consequence of the GUM assumption about the measurement process.

c) Let $\bar{\theta}$ be the deviation of the average temperature of the test bed from the nominal value of 20 °C. An estimate of $\bar{\theta}$ is –0,1 °C with a standard deviation of 0,2 °C. Since the GUM gives no additional information concerning this standard deviation, infinite degrees of freedom is assumed for it and also that $\bar{\theta}$ is a draw from a Gaussian distribution. Hence

$$\tilde{\bar{\theta}} = -0,1 - 0,2Z, \quad (41)$$

where Z is a draw from a standard Gaussian random variable, independent of all other random variables.

d) An FQ for Δ has a probability density function given by

$$g(x) = \frac{2}{\pi\sqrt{1-4x^2}}, \quad -0,5\ ^\circ\text{C} < x < 0,5\ ^\circ\text{C}. \quad (42)$$

For making random draws from the arcsine distribution (42), observe that if U_1 is a uniform (0,1) random variable, $-\cos(\pi U_1)/2$ has the required arcsine distribution. So, an FQ for Δ may be taken to be

$$\tilde{\Delta} = -\cos(\pi U_1)/2. \quad (43)$$

e) An FQ for δ_α is

$$\tilde{\delta}_\alpha = U_2, \quad (44)$$

where U_2 is a uniform random variable over the interval ±1 × 10^{-6} °C^{-1}.

f) An FQ for δ_θ is

$$\tilde{\delta}_\theta = U_3, \tag{45}$$

where U_3 is a uniform random variable over the interval ±0,05 °C.

g) An FQ for α_s is

$$\tilde{\alpha}_s = 11{,}5 \times 10^{-6} + U_4, \tag{46}$$

where U_4 is a uniform random variable over the interval ±2 × 10^{-6} °C^{-1}.

11.4.2 Substituting the fiducial quantities in Expressions (37) to (46) into Expression (31), a fiducial quantity is obtained for λ. An approximation to the distribution for $\tilde{\lambda}$ is provided using 500 000 Monte Carlo trials. The mean and the standard deviation of this approximate distribution are 50 000 838 nm and 35 nm, respectively. A 95 % fiducial interval for λ is given by the 0,025 and 0,975 quantiles of this distribution, namely, (50 000 768 nm 50 000 907 nm). R code for generating the 500 000 realizations of $\tilde{\lambda}$ follows.

```
nrun = 500000
lambda.s = 50000623 - 25 * rt(nrun, 18)
delta.lambda = 215 - 13/sqrt(5) * rt(nrun, 24)
delta.cr = 3.9*rt(nrun, 5)
delta.cnr = 6.7*rt(nrun, 8)
theta = rnorm(nrun, -0.1, 0.2)
delta = (-cos(pi*runif(nrun))/2)
delta.alpha=runif(nrun,-10^(-6), 10^(-6))
delta.theta=runif(nrun, -0.05, 0.05)
alpha.s=runif(nrun, (11.5-2)*10^(-6), (11.5+2)*10^(-6))
lambda=(lambda.s*(1 + alpha.s*(theta + delta - delta.theta))+delta.lambda
         +delta.cr + delta.cnr)/(1 + (alpha.s + delta.aplha)*(theta+delta))
```

12 Discussion

12.1 Comparison of uncertainty evaluations using the three statistical approaches

12.1.1 Table 4 summarizes the results for Example 1. The frequentist bootstrap, Bayesian and fiducial solutions for Example 1a and Example 1b are very similar. The bootstrap and the GUM solutions produce slightly shorter intervals in both Example 1a and Example 1b. More substantial differences are seen in the solution for Example 1c. Here the Bayes' solution based on the uniform prior density produces an interval that is markedly longer than most of the other methods; only the conservative Eisenhart interval is wider.

Table 4. Expanded uncertainty intervals for the three statistical approaches for Example 1

	GUM	Eisenhart	Bootstrap	Bayes	Fiducial
Example 1a	(1,89 2,73)	(1,89 2,73)	(1,83 2,66)	(1,81 2,82)	(1,86 2,76)
Example 1b	(1,90 2,72)	(1,78 2,84)	(1,86 2,64)	(1,83 2,79)	(1,87 2,75)
Example 1c	(0,00 0,12)	(0,00 0,20)	(0,00 0,11)	(0,00 0,19)	(0,00 0,14)

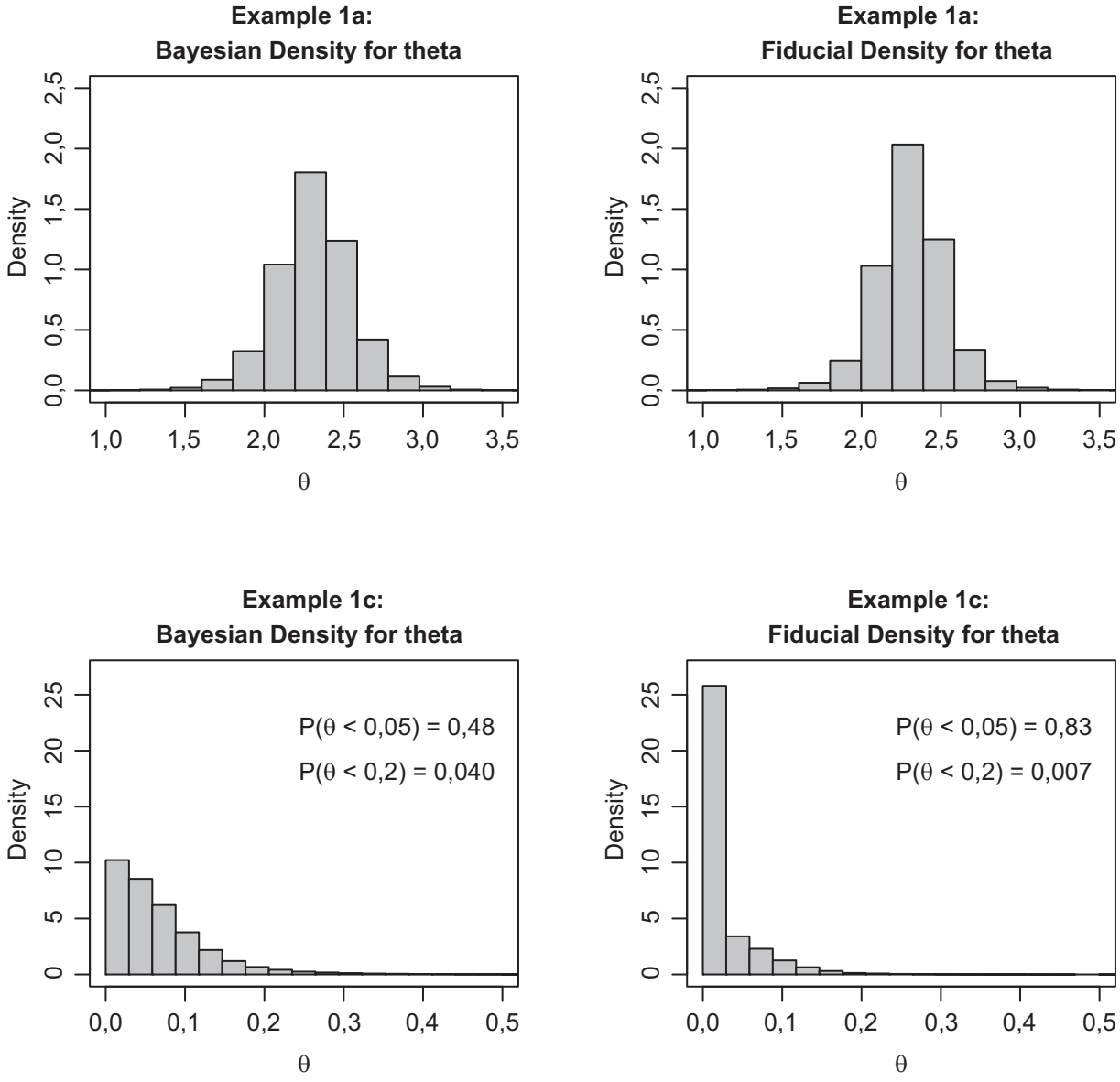

Figure 2 — Comparison of the approximate Bayesian and fiducial densities for Examples 1a and 1c

12.1.2 Because the Bayesian and fiducial approaches produce probability distributions for the measurand, θ, their results for Example 1a and Example 1c are further compared in Figure 2, in addition to the comparison of the expanded uncertainty intervals in Table 4. The results for Example 1b are not displayed because they are visually indistinguishable from those for Example 1a. From the histograms in Figure 2 it is clear that the Bayesian posterior probability distribution for θ and the fiducial distribution for θ are similar when the signal appreciably exceeds the background. When the signal is near the background, however, the two distributions have very different characteristics due to their different methods of incorporating the physical constraints inherent in the problem.

12.1.3 In the frequentist context, the measurand θ and the input quantities $\mu_1,...,\mu_p$ in the measurement model (1) are all assumed to be fixed unknown constants. This approach seems to be quite reasonable if the measurand represents a physical constant for which previous studies do not provide an appropriate (informative) prior distribution or a structural equation. It is favoured by statisticians who do not believe that all parameters must be modelled as random variables, although it typically handles uncertainties obtained by Type B methods of evaluation by assigning them a probability distribution and integrating over this distribution. In this regard, it is similar to the Bayesian approach, where all parameters are characterized by probability distributions, but it needs fewer distributional assumptions.

12.1.4 The bootstrap is a well-established statistical method that can replace complicated and often inaccurate approximate confidence intervals by computer simulations. There are various bootstrap schemes developed to construct confidence intervals under different conditions. The parametric bootstrap-*t* interval, introduced in this document, is the natural choice as an improvement to the Student-*t* interval of the GUM. The advantage of bootstrapping is its simplicity – it is straightforward to apply the bootstrap to derive confidence intervals as demonstrated in the examples.

12.1.5 It was shown with the examples that Bayesian uncertainty evaluation using a statistical model is conceptually simple, and can be applied to complicated measurement problems without any change to the basic method. Systematic effects, which cannot be estimated from measured values (that is, there are no functions of the observations whose expected values are equal to the systematic effect) and for which information is used to perform a Type B evaluation of uncertainty, can easily be included in the Bayesian model. Computation of posterior distributions can be carried out using MCMC methods, often using existing software. As was seen, there is no need for asymptotic arguments to justify the probability statements since small and large samples share a common, probabilistic justification.

12.1.6 There are some drawbacks to the Bayesian methods described here. Most importantly, to use them prior distributions are to be specified for all parameters in the measurement model, including the measurand. Even though in metrology informative prior distributions are often available in the form of Type B uncertainty evaluations, it is usually the case that one or two of the parameters will need to be assigned vague (non-informative) prior distributions because of lack of prior knowledge. Such distributions are not unique, and as was demonstrated in Example 1c, they can influence the results. It is therefore advisable to perform sensitivity analyses to judge the magnitude of such effects. Large effects arising from the specification of a non-informative prior require further study of the measurement system. Generally the presence of such effects means that there is insufficient information available about the measurand in the data and thus the prior distribution has considerable influence on the result. In some cases, this problem can be addressed by increasing the number of replicated measured values, or by changing the way in which the data is collected, for example by improving resolution. In other situations, it may be that the mathematical model being used has too many parameters about which no real prior information is known and so the model should be simplified.

12.1.7 When substantial prior information about the measurand does exist, it can be introduced simply, and updated efficiently via Bayes' theorem. Further, sensitivity to the prior form, not just for the measurand, but also for the standard deviation of the likelihood function, can be a good indication that there are problems with the measurement system. These can then be studied and corrected.

12.1.8 Fiducial inference provides a framework for associating a distribution with a parameter of interest. Recent research results [26] show that fiducial inference is a valid statistical method with generally good operating characteristics. The examples used demonstrated that the fiducial approach could easily and naturally incorporate the uncertainty information into the measurement model, and obtain the estimate of the measurand and its associated standard uncertainty by propagating the component statistical distributions.

There is no need for propagation of uncertainty based on Taylor series expansions or the Welch-Satterthwaite approximation under the fiducial approach.

12.1.9 There is an issue of non-uniqueness in the fiducial distribution due to the choice of a particular form of the structural equation. However, it is important to note that, in many practical applications, the physical process by which the data was generated is known. In this case, the structural equation should be chosen to reflect this process, thus eliminating the problem of non-uniqueness. In metrology, where an unknown measurand is measured using some known processes, it is known that random errors influence the measurement in some specified fashion. The resulting measured values are expressed in terms of a measurement model that combines the measured quantities and errors in the form of influence quantities. This model can be taken as the structural equation.

12.2 Relation between the methods proposed in GUM Supplement 1 (GUMS1) and the three statistical approaches

12.2.1 GUMS1[3] generates random draws from a probability distribution for an output quantity Y in a measurement model that "describes the knowledge of that quantity, based on the knowledge of the input quantities, as described by the PDFs assigned to them" (page vii of Reference [3]). On the same page, GUMS1 states that the "PDF for a quantity is not to be understood as a frequency density." Finally, on page 7, it defines the output quantity Y to be the measurand. Thus, the results of the GUMS1 analysis, such as the mean and standard deviation of the Monte Carlo draws, are estimates of features of a probability density for the measurand. Accordingly, a direct comparison between the results from GUMS1 and the fiducial method as well as the traditional Bayesian method, is possible. The GUMS1 uncertainty intervals may be studied for their frequentist coverage properties but they should not be interpreted as usual frequentist confidence intervals.

12.2.2 As indicated in 9.1.1 and 9.1.2, traditional Bayesian methods are based on a statistical model that accommodates prior knowledge of the measurand. This statement is not true of GUMS1 because this method is based on a measurement model, where the measurand is the output quantity, and so its probability distribution is completely determined by the probability densities of the input quantities. (This is also stated on pages 2 and 8 of Reference [3].) Thus, any direct comparison of results from traditional Bayesian methods and GUMS1 methods are limited to the case of no prior knowledge of the measurand.

12.2.3 Reference [34] performs such a comparison for a particular but widely applicable measurement problem. In this paper, the measurand μ is a function of α and β, that is, the measurement model is $\mu = f(\alpha, \beta)$. The parameter α can be estimated from data because it is the mean of a Gaussian random variable X, for which there is a set of observed values. No data is available to estimate the parameter β, but a belief distribution is given. The GUMS1 analysis (see 6.4.9.2 of Reference [3]) assigns a scaled and shifted Student-*t* distribution to α and then propagates the distributions for α and β using the function f. In Reference [34], it is shown that this analysis is equivalent to a Bayesian calculation of the probability density for the function $f(\alpha, \beta)$, where the two parameters are taken to be independent, the likelihood function for X is Gaussian with mean α, an improper uniform prior distribution is used for α, and the density for β is given by the belief distribution. Note that a prior density for μ is not used here.

12.2.4 Suppose now that there exists a function g such that $\alpha = g(\mu, \beta)$. Traditional Bayesian analysis uses a Gaussian likelihood function for X with mean $g(\mu, \beta)$ and prior distributions for μ and β. In the absence of any additional information about the measurand, the improper uniform distribution might be used for μ, but there are also other choices. The belief distribution of β is a natural choice for the prior. Usually, μ and β would be taken as independent random variables. Note that for this model, a prior density for α is not used.

12.2.5 GUMS1 and traditional Bayesian analysis use different parametrizations of the same statistical model. Ignorance of the measurand is expressed differently under each parametrization. The model used by GUMS1 does not do this directly by means of a density for μ, but instead uses a non-informative, improper prior for α. As indicated in Reference [34], this is a different assumption. Traditional Bayesian analysis generally uses a non-informative prior for the measurand μ, itself. Reference [34] shows that the two analyses yield identical probability distributions for the measurand when the improper uniform prior for μ is used in the

Bayesian analysis, and the function f is linear. For non-linear functions, the probability distributions for μ derived under the two parametrizations are not the same. It is important to note that if the non-informative prior for α is transformed into a prior for μ then the corresponding traditional Bayes analysis yields the same results as GUMS1 for any function.

12.2.6 As stated earlier, based on the measurement model, GUMS1 obtains a PDF for the measurand by propagating the PDFs for the input quantities. The resulting PDF describes knowledge of the measurand given the observed data and assumptions made in assigning the joint PDF for the input. In many standard models with measured values regarded as draws from univariate normal distributions, uncertainty intervals obtained using GUMS1 and fiducial methods are very similar, if not identical. Recall the measurement model in Example 1a, namely

$$\theta = \gamma - \beta$$

with $Y_i \sim N(\gamma, \sigma_Y^2)$, $i = 1,...,5$ and $B_j \sim N(\beta, \sigma_B^2)$, $j = 1,...,5$. Based on the guidance in GUMS1, a scaled and shifted *t*-distribution is assigned as the PDF for γ and the PDF for β. Specifically, the PDF for γ has the same distribution as the random variable

$$\bar{y} - \frac{s_y}{\sqrt{5}} T_4^{(1)},$$

where $T_4^{(1)}$ is a Student's-*t* random variable with 4 degrees of freedom, and the PDF for β has the same distribution as the random variable

$$\bar{b} - \frac{s_b}{\sqrt{5}} T_4^{(2)},$$

where $T_4^{(2)}$ is a Student's-*t* random variable with 4 degrees of freedom that is independent of $T_4^{(1)}$. Consequently, the PDF for the measurand θ can be obtained from the distribution for

$$\bar{y} - \bar{b} - \frac{s_y}{\sqrt{5}} T_4^{(1)} + \frac{s_b}{\sqrt{5}} T_4^{(2)}.$$

R code for generating the 500 000 realizations of the above distribution is listed below.

```
nrun = 500000
T1 = rt(nrun, 4)
T2 = rt(nrun, 4)
theta = 3.537 - 1.228 - 0.342/sqrt(5)*T1 + 0.131/sqrt(5)*T2
```

A 95 % uncertainty interval based on the 0,025 and 0,975 quantiles of the approximate PDF is

```
quantile(theta, c(0.025, 0.975))
## 2.5%  97.5%
## 1.853703 2.763999
```

which is essentially identical to the fiducial interval for this example given earlier. Similarly, the GUMS1 and fiducial approaches produce the same uncertainty interval for problems in Examples 1b and 1c.

12.2.7 In many other cases, the GUMS1 and fiducial methods produce different results. An “extreme” case can be found in a problem described in Reference [35]. In the example presented in Reference [35], the measurand is the magnitude of a complex-valued quantity.

$$\Gamma = \Gamma_1 + i\Gamma_2.$$

That is, the measurand is

$$|\Gamma| = \sqrt{\Gamma_1^2 + \Gamma_2^2}.$$

Assuming $X_1 \sim N(\Gamma_1, \sigma^2)$ and $X_2 \sim N(\Gamma_2, \sigma^2)$ with known σ, GUMS1 assigns $N(x_1, \sigma^2)$ to the PDF for Γ_1 and $N(x_2, \sigma^2)$ to the PDF for Γ_2. Consequently, the PDF for $|\Gamma|$ based on GUMS1 is the distribution of for random variable

$$\sqrt{(x_1 - \sigma Z_1)^2 + (x_2 - \sigma Z_2)^2}, \tag{43}$$

where Z_1 and Z_2 are independent standard normal random variables. Reference [35] showed that the GUMS1 intervals for $|\Gamma|$ have unsatisfactory frequentist performance (insufficient coverage probabilities) when $|\Gamma|$ is small compared to σ. This is because the random variable in Expression (43) is positive and hence the lower bound of the uncertainty interval for $|\Gamma|$ will be positive also, which may not cover $|\Gamma|$ when $|\Gamma|$ is close to 0.

12.2.8 A fiducial solution for the problem can be derived based on the fact that $(X_1^2 + X_2^2)/\sigma^2$ is distributed as a non-central χ^2 with 2 degrees of freedom and non-centrality parameter $\lambda = |\Gamma|^2/\sigma^2$. This distributional property can be used to develop a structural equation that relates the observable statistic $(X_1^2 + X_2^2)/\sigma^2$ to λ, which contains the parameter of interest $|\Gamma|$. Based on this structural equation, a fiducial interval for $|\Gamma|$ can be constructed. Reference [36] showed that this fiducial interval maintains the nominal frequentist coverage in all situations.

13 Summary

13.1 In this Technical Report, three approaches for constructing uncertainty intervals that have clear probabilistic interpretations are discussed. In contrast, much other work in this area has focused on assessing the statistical properties of procedures currently in popular use across the metrology community. One of the goals in approaching the study of methods for uncertainty evaluation from such a vantage point was to try to gain insight into current methods and highlight new options that may also prove useful.

13.2 As Reference [9] observed, the uncertainty intervals obtained under the different approaches will often be similar numerically. Even when this is case, however, their interpretations are different.

13.3 Frequentist uncertainty intervals make probabilistic statements about the long-term performance of a particular procedure for constructing uncertainty intervals during repeated use under identical conditions. Thus, the probability statement is not directly about the value of the measurand, but is about the long-term relation between the procedure by which the interval has been constructed and the measurand. Once measured values have been obtained and a frequentist uncertainty interval has been computed, there is no longer anything random about the results. Although it is not known whether the value of the measurand is captured in any particular interval, such intervals will capture the value of the measurand with a specified probability. Unlike a traditional confidence interval based only on statistical data, the frequentist uncertainty interval is typically constructed so that the desired confidence level is attained on average after integrating over the probability distributions of any quantities that are obtained using Type B evaluations of uncertainty.

13.4 Bayesian and fiducial uncertainty intervals, on the other hand, are based on probability distributions that directly describe knowledge of the value of the measurand. The methods used to obtain these two types of intervals are different, but the results are similar in this aspect of their interpretation. The Bayesian results are obtained by combining probability distributions for each parameter specified prior to analysis of the data with a probability model that describes the variation in the data using Bayes' theorem. The resulting posterior distributions for each parameter reflect the probability of the parameter values given the prior information and

the data. The fiducial results are obtained by inverting a probability model for the data given the parameters to obtain a distribution for the parameter values given the data.

13.5 If the numerical results were always similar, each of the interpretations would be applicable (at least approximately) to every uncertainty interval. However, as demonstrated in the examples in this Technical Report, the numerical results can differ appreciably from one another in some instances, even though each can be justified probabilistically and they share a common level of significance (generally 95 %). Other differences also may be observed. For example, if one of the dominant sources of uncertainty in a particular application corresponds to a quantity having a skewed distribution, the uncertainty intervals obtained using the Bayesian or fiducial approaches reflect that asymmetry, while an approximate confidence interval obtained using the procedures of the GUM will produce a symmetric uncertainty interval (and may be longer than necessary on one side). Frequentist results based on other statistical principles may match the Bayesian or fiducial results in some cases, but the different methods will not agree in general because each approach is ultimately based on a different set of mathematical assumptions and criteria.

13.6 The existence of different approaches for uncertainty evaluation that do not always agree might be seen as a complication. However, it is better seen as an indication of further opportunity. It is only by continually working together to appreciate the features of different approaches that methods for uncertainty evaluation will ultimately be obtained that meet all the great bulk of scientific and economic needs: methods that are practical to implement, make efficient use of resources, are applicable to many types of measurement, both old and new, and are transparent in meaning.

Bibliography

[1] ISO/IEC Guide 98-3:2008, *Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)*

[2] ISO/IEC 17025:2005, *General requirements for the competence of testing and calibration laboratories*, 2005

[3] ISO/IEC Guide 98-3:2008/Suppl 1:2008, *Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995) — Supplement 1: Propagation of distributions using a Monte Carlo method*

[4] ISO/IEC Guide 99:2007, *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*

[5] GLESER, L.J. Assessing uncertainty in measurement. *Statistical Science*, 13:277–290, 1998

[6] KACKER, R. and JONES, A. On use of Bayesian statistics to make the guide to the expression of uncertainty in measurement consistent. *Metrologia*, 40:235–248, 2003

[7] ELSTER, C. W., WÖGER, W. and COX, M.G. Draft GUM Supplement 1 and Bayesian analysis. *Metrologia*, 44:L31–L32, 2007

[8] WILLINK, R. A procedure for the evaluation of measurement uncertainty based on moments. *Metrologia*, 42:329–343, 2005

[9] LIRA, I. and WÖGER, W. Comparison between the conventional and Bayesian approaches to evaluate measurement data. *Metrologia*, 43: S249–S259, 2006

[10] GUTHRIE, W.F., LIU, H.K., RUKHIN, A.L., TOMAN, B., WANG, C.M. and ZHANG, N.F. Three statistical paradigms for the assessment and interpretation of measurement uncertainty. *Data Modeling for Metrology and Testing in Measurement Sciences*, edit. Pavese, F. and Forbes, A. B. Birkhauser, Boston, 2008

[11] KIRKUP, L. and FRENKEL, B. *An Introduction to Uncertainty in Measurement Using the GUM*. Cambridge University Press, Cambridge UK, 2006

[12] BAYES, T. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1764. (facsimile available at http://www.stat.ucla.edu/history/essay.pdf)

[13] FISHER, R.A. Inverse probability. *Proc. Comb. Philos. Soc.*, 26:528–535, 1930

[14] CASELLA, G. and BERGER, R. *Statistical Inference*. Duxbury, MA, 2 edition, 2002

[15] GUTHRIE, W.F. Should (T1-T2) have larger uncertainty than T1? *Proceedings of the 8th International Conference on Temperature: Its Measurements and Control*, 2:887–892, 2002. http://www.itl.nist.gov/div898/pubs/author/guthrie/guthrie-2002-01.pdf)

[16] EFRON, B. and TIBSHIRANI, R.J. An Introduction to the Bootstrap. *Monographs of Statistics and Applied Probability*, volume 57. Chapman and Hall, 1993

[17] EISENHART, C. Expression of the uncertainties of final measurements results. *NBS special publication*, NIST, Gaithersburg, MD, 1983

[18] R Development Core Team. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, 2003. ISBN 3-900051-00-3, http://www.R-project.org

[19] LUNN, D.J, THOMAS, A., BEST, N. and SPIEGELHALTER, D. WinBUGS – a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325-337, 2000

[20] BERNARDO, J.M. and SMITH, A.F.M. *Bayesian Theory*. John Wiley and Sons Ltd, 1994

[21] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001

[22] GELMAN, A., CARLIN, J.B., STERN, H.S., and RUBIN, D.B. *Bayesian Data Analysis*. Chapman and Hall, 1995

[23] BROWNE, W.J. and DRAPER, D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1:473–514, 2006

[24] WANG, C.M. and IYER, H.K. Propagation of uncertainties in measurements using generalized inference. *Metrologia*, 42:145–153, 2005

[25] WANG, C.M. and IYER, H.K. A generalized confidence interval for a measurand in the presence of Type-A and Type-B uncertainties. *Measurement*, 39:856–863, 2006

[26] HANNIG, J., IYER, H. K., and PATTERSON, P. L. Fiducial generalized confidence intervals. *Journal of the American Statistical Association*, 101:254–269, 2006

[27] WANG, C.M. and IYER, H.K. Uncertainty analysis for vector measurands using fiducial inference. *Metrologia*, 43:486–494, 2006

[28] FRASER, D.A.S. *The Structure of Inference*. New York: Krieger, 1968

[29] HANNIG, J. On fiducial inference – the good, the bad and the ugly. Technical Report 2006/3, Department of Statistics, Colorado State University, Fort Collins, CO, 2006. URL http://www.stat.colostate.edu/research/2006_3.pdf

[30] IYER, H.K. and PATTERSON, P.L. A recipe for constructing generalized pivot quantities and generalized confidence intervals. Technical Report 2002/10, Department of Statistics, Colorado State University, Fort Collins, CO, 2002. URL http://www.stat.colostate.edu/research/2002_10.pdf

[31] RUKHIN, A.L. and SEDRANSK, N. Statistics in metrology: international key comparisons and interlaboratory studies. *Journal of Data Science*, 7:393–412, 2007

[32] EFRON, B. Six questions raised by the bootstrap. In: *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, editors) pages 99–126. Wiley, NY, 1992

[33] LINDLEY, D. and SMITH, A.F.M. Bayes estimates for the linear model, *JRSS B*., 34:1-41, 1972

[34] ELSTER, C. and TOMAN, B. Bayesian uncertainty analysis under prior ignorance of the measurand versus analysis using the Supplement 1 to the Guide: a comparison, *Metrologia*, 46:261-266, 2009

[35] HALL, B.D. Evaluating methods of calculating measurement uncertainty, *Metrologia*, 45:L5-L8, 2008

[36] WANG, C.M. and IYER, H.K. Fiducial intervals for the magnitude of a complex-valued quantity, *Metrologia*, 46:1 81-86, 2009

ICS 03.120.30

Price based on 43 pages