A brief guide to measurement uncertainty (IUPAC Technical Report)

Possolo, Antonio; Hibbert, David Brynn; Stohner, Jürgen; Bodnar, Olha; Meija, Juris

Pure and Applied Chemistry, 2024-01-18

IUPAC Technical Report

Antonio Possolo*, David Brynn Hibbert, Jürgen Stohner, Olha Bodnar and Juris Meija

A brief guide to measurement uncertainty (IUPAC Technical Report)

https://doi.org/10.1515/pac-2022-1203
Received December 16, 2022; accepted November 10, 2023

Abstract: This *Brief Guide* reintroduces readers to the main concepts and technical tools used for the evaluation and expression of measurement uncertainty, including both classical and Bayesian statistical methods. The general approach is the same that was adopted by the *Guide to the Expression of Uncertainty in Measurement* (GUM): quantities whose values are surrounded by uncertainty are modeled as random variables, which enables the application of a wide range of techniques from probability and statistics to the evaluation of measurement uncertainty. All the methods presented are illustrated with examples involving real measurement results from a wide range of fields of chemistry and related sciences, ranging from classical analytical chemistry as practiced at the beginning to the 20th century, to contemporary studies of isotopic compositions of the elements and clinical trials. The supplementary material offers profusely annotated computer codes that allow the readers to reproduce all the calculations underlying the results presented in the examples.

Keywords: Bayesian methods; Gauss's formula; measurement uncertainty; Monte Carlo methods; uncertainty propagation.

CONTENTS

Sponsoring body: IUPAC Analytical Chemistry Division Committee: see more details on page 17.

***Corresponding author: Antonio Possolo**, National Institute of Standards and Technology, Gaithersburg, MD, USA; and Georgetown University, Washington, DC, USA, e-mail: antonio.possolo@nist.gov. https://orcid.org/0000-0002-8691-4190
David Brynn Hibbert, School of Chemistry, UNSW, Sydney, NSW 2052, Australia. https://orcid.org/0000-0001-9210-2941
Jürgen Stohner, Institute of Chemistry and Biotechnology, Zürich University of Applied Sciences, Einsiedlerstrasse 31, CH-8820 Wädenswil, Switzerland. https://orcid.org/0000-0003-0079-7083
Olha Bodnar, Örebro University School of Business, Örebro, Sweden; and National Institute of Standards and Technology, Gaithersburg, MD, USA. https://orcid.org/0000-0003-1359-3311
Juris Meija, National Research Council Canada, Ottawa, ON, Canada. https://orcid.org/0000-0002-3349-5535

1 Introduction

IUPAC has been contributing to the development of the *Guide to the Expression of Uncertainty in Measurement* since 1997, including its original version [1] and related documents, which are collectively known as the GUM. Since these documents make extensive use of rather specialized concepts and methods of probability and statistics, many users and producers of evaluations of measurement uncertainty resort to less technical, more accessible guidance and varied illustrative examples such as are provided by the EURACHEM Guide for *Quantifying Uncertainty in Analytical Measurement* [2] or by the *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results* [3].

The task of evaluating measurement uncertainty is often reduced to calculating standard deviations, without careful consideration of the measurement models that explain the interplay between the contributions made by the relevant sources of uncertainty. The assumptions that validate such models and that make the uncertainty evaluations fit for purpose often remain unexamined. In many applications, relative uncertainty contributions are incorrectly combined into a sum of their squares when the underlying measurement model is not stated, or even when the measurand is not a product of powers of input quantities (see Subsection 4.2).

This *Brief Guide* provides a concise reintroduction to models and methods for uncertainty evaluation that emphasizes how such evaluation depends on the purpose that the measurement results are intended for and includes practical illustrations involving real data that are representative of the range of problems likely to be encountered in chemical science and engineering.

1.1 Notation

When referring to a set of values of one or several quantities, $\{x_1,\ldots,x_n\}$, in this contribution, we often represent this set as $\{x_i\}$ for brevity, to distinguish between the entire set and a single element x_i, consistent with established mathematical notation. Vectors, matrices, and more generally ordered sets are represented by symbols in boldface. The symbol $[x_1 \ \ldots \ x_m]$ refers to a row vector, which is a matrix with 1 row and m columns, whereas $[x_1 \ \ldots \ x_m]^\top$ denotes its transpose: a column vector or matrix with m rows and 1 column. Similarly, if $\boldsymbol{J}$ denotes a matrix with m rows and n columns, then $\boldsymbol{J}^\top$ denotes its transpose, a matrix with n rows and m columns, whose columns are the rows of $\boldsymbol{J}$. Although often omitted for brevity, the circumflex (hat) symbol signifies the estimate of the quantity. For example, the maximum likelihood estimate of a mass, m, is denoted as $\hat{m}$. The overline, $\overline{m}$, is used to denote the (equally weighted) arithmetic average of replicated determinations of m.

Example 3E introduces the parenthetic notation to represent standard uncertainties where the estimate of the molar mass of iridium is shown as 192.216 62(29) g/mol. This denotes a measured value of 192.216 62 g/mol and an associated standard uncertainty of 0.000 29 g/mol. This parenthetic notation is used extensively throughout this *Brief Guide* and more generally in many fields of science and technology. The number between parentheses denotes the standard uncertainty, which affects as many trailing (least significant) digits of the measured value as there are digits between the parentheses, regardless of the position of the decimal point, if any.

2 Measurement

Measurement is an experimental or computational process that, by comparison with a standard, produces an estimate of the true value of a (quantitative or qualitative) property of a material, virtual object, collection of

objects, process, or event or a series of events [4]. A *measurement result* is an estimate of the property intended to be measured (*measurand*) qualified with an evaluation of the associated measurement uncertainty. Measurements of some kinds of nominal properties are often referred to as *examinations*, *identifications*, or *classifications*.

Example 2A: (Quantitative and qualitative measurands) Since measurands can be quantitative or qualitative, the corresponding measurement results reflect this difference. Quantitative measurands have values for which differences or ratios are meaningful. Neither of these arithmetic operations are meaningful for qualitative (or, categorical) measurands, which can be further classified according to whether they are ordinal or nominal: values of the former can be ordered, from smallest to largest, or from weakest to strongest; values of the latter can be compared only by equality (whether they are the same or different).

QUANTITATIVE: The certificate of analysis for drinking water certified reference material AQUA-1 [5], developed by the National Research Council Canada, lists 0.222 µg/L as certified value for the mass concentration of arsenic, with an associated standard uncertainty of 0.007 µg/L.

ORDINAL: Mineralab (Eagle, Idaho) produces a hardness test kit, HD-2, for industrial use, including eight picks that deliver values from 2 to 9 of Mohs' hardness [6, §3.5] for minerals and other natural and manufactured products.

NOMINAL: The National Institute of Standards and Technology (NIST, U.S.A.) is most confident in the identity of Reference Material 8256, Wild-caught Coho Salmon, as consisting of tissue from *Oncorhynchus kisutch* [7]. The identification involved sequencing DNA fragments and comparing them against reference, identifying sequences of voucher specimens of this species in the Reference Standard Sequence Library of the U.S. Food and Drug Administration (FDA).

Measurement aims "*to find* out something" [8] and is made to support decision-making [9, 10]. For example: the measurement of the concentration of cardiac-specific troponins in the blood enables decisions about how to manage myocardial infarctions; the measurement of the pH of the water in a swimming pool informs the decision regarding its maintenance; and the measurement of the internal temperature of cooked meat determines whether it is ready and safe to eat.

Measurement uncertainty is the doubt about the true value of the measurand that remains after making a measurement [4, 11]. It can be described fully and quantitatively by a probability distribution on the set of values of the measurand, or, more concisely for scalar measurands, by the standard deviation of such distribution (which is called the *standard uncertainty*), or, for multivariate measurands, by its covariance matrix (also called *dispersion matrix*).

If y denotes a scalar measurand, then we write $u(y)$ to denote the standard uncertainty associated with y, mindful of the fact that the uncertainty is about the true value of the measurand, not about estimates of it. In cases where different estimates of the same measurand are under consideration, as in Example 4F, and one needs to refer to the different uncertainty evaluations, one can refer to them as $u_1(y)$ and $u_2(y)$, for example.

It should be noted that the meaning of $u(y)$ depends on the probability distribution that is used to model the uncertainty associated with y. For example, the probability that the interval $y \pm u(y)$ covers the true value of the measurand is 68 % when y has a Gaussian (or, normal) distribution, 58 % when y has a uniform (or, rectangular) distribution, and 76 % when y has a Laplace (or, double exponential) distribution.

3 Measurement models

Measurement models relate the measurand (output) to inputs whose values are either observable during the current experiment or have been measured previously [12]. We assume throughout that both the output and the inputs are quantities (that is, properties whose values can be compared by ratios or by differences). However, measurement models generally can incorporate qualitative inputs as well.

3.1 Deterministic measurement models

Deterministic measurement models are functional relationships between the true value of the output η and the true values of n input quantities, $\{\xi_1,\ldots,\xi_n\}$, expressed as $\eta = f(\xi_1,\ldots,\xi_n)$, where the *measurement function* f is determined by a physical theory, empirically, or by convention. Note that the very definition of f can be surrounded by uncertainty, which should be recognized and propagated whenever it makes a non-negligible contribution to the uncertainty associated with the estimate of η.

Example 3A: (Theoretical model) The amount concentration of sodium chloride, c(NaCl), in an aqueous solution of volume V(NaCl) is determined by titration with a silver nitrate solution in the presence of K_2CrO_4. The measurement function is derived from the chemical theory that explains the reaction between NaCl and $AgNO_3$,

$$NaCl + AgNO_3 \rightarrow NaNO_3 + AgCl\downarrow,$$

and the endpoint of this titration is determined by the appearance of the brown-red silver chromate. A simple theoretical measurement model assumes that Ag_2CrO_4, being more soluble than AgCl, will start forming only after the chloride ions will have been consumed by the silver ions. This leads to the familiar relationship $c(NaCl)V(NaCl) = c(AgNO_3)V(AgNO_3)$, where $V(AgNO_3)$ is the volume of the silver nitrate solution that was used to reach the endpoint, hence the measurement model $c(NaCl) = c(AgNO_3)V(AgNO_3)/V(NaCl)$ [13].

A more refined physico-chemical measurement model considers that AgCl and Ag_2CrO_4 are somewhat soluble, and the silver chromate will start to precipitate when $c^2(Ag^+) \times c(CrO_4^{2-}) > K_{sp}(Ag_2CrO_4)$, where K_{sp} denotes the solubility product. Depending on the concentration of potassium chromate, this can happen either slightly before or after the equivalence point is reached [14].

Example 3B: (Empirical model) In the mass spectrometric analysis of biomolecules, the signal, A, produced by the mass spectrometer is often not directly proportional to the amount concentration, c, of these biomolecules. If the reasons for these deviations cannot be fully understood, one often uses empirically adequate mathematical functions to model the approximate relationship between the concentration and the analytical signal [15], for example, the quadratic function $A = a_0 + a_1c + a_2c^2$.

Example 3C: (Conventional model) The total hardness of water, H, is often defined as the mass concentration of calcium carbonate that is equivalent to the sum of the mass concentrations of dissolved calcium and magnesium, $H = \gamma(Ca^{2+}) \times M(CaCO_3)/M(Ca) + \gamma(Mg^{2+}) \times M(CaCO_3)/M(Mg)$. The specific measurement function that relates the output quantity, H, to the measured input quantities, $\gamma(Ca^{2+})$ and $\gamma(Mg^{2+})$, is defined by convention. Another example is the anion gap, c_{AG}, defined as the difference between selected cation and anion amount concentrations in urine, serum, or plasma: $c_{AG} = c(K^+) + c(Na^+) - c(Cl^-) - c(HCO_3^-)$ [16].

3.2 Statistical measurement models

Measurement models often comprise deterministic and stochastic components. The stochastic components invoke probability distributions to model the uncertainties associated with inputs and output. This is tantamount to regarding these quantities as *random variables*, which we define intuitively as quantities that have probability distribution as attributes, and where the adjective "random" does not suggest that the quantity's true value varies unpredictably, as if by chance – it simply indicates that the quantity's true value is not known with full certainty. Accordingly, the relationship between the observed input quantities is rewritten as an errors-in-variables regression [17] $y = f(x_1,\ldots,x_n)$, where f is the measurement function, $y = \eta + \Delta y$ and $x_i = \xi_i + \Delta x_i$, with Δy and Δx_i denoting non-observable measurement errors, ξ_i denoting the true value of x_i, and η denoting the true value of y, for $i = 1,\ldots,n$.

Example 3D: (Measurement error model) Consider n replicate determinations of the mass, $m_{\text{obs},1},\ldots,m_{\text{obs},n}$, and of the volume, $V_{\text{obs},1},\ldots,V_{\text{obs},n}$, of a liquid, made to estimate its density, ρ. The deterministic measurement model for the density is $\rho = m/V$ which relates the true values (that is, free of measurement errors) of these quantities. The observed values of the mass and volume are related to their true values as follows:

$$m_{\text{obs},i} = m + \Delta m_i \quad \text{and} \quad V_{\text{obs},i} = V + \Delta V_i,$$

where Δm_i and ΔV_i are measurement errors modeled as outcomes of independent Gaussian random variables, all with mean 0, and standard deviations equal to their associated, reported standard uncertainties, $u(m_{\text{obs},i})$ and $u(V_{\text{obs},i})$, respectively, for $i = 1,\ldots,n$.

The deterministic component of a statistical measurement model expresses the measurand, $\eta = g(\theta)$, as a specified function g of a parameter θ (which, in Section 4.4, for example, is a vector with several components) that indexes the joint probability distribution of the input quantities, which characterizes the stochastic component of the model. The value of θ has to be estimated, typically from replicated observations of the inputs, to obtain an estimate of the measurand as $\widehat{\eta} = g(\widehat{\theta})$. The method of moments, least squares, maximum likelihood estimation (4.4), and Bayesian methods (4.7) are commonly used to obtain $\widehat{\theta}$ [18]. These different methods typically produce not only different estimates of θ but also different evaluations of $u(\theta)$.

Example 3E: (Molar mass of iridium – laboratory effects model) Prohaska et al. [19, §8.4] based the calculation of the 2021 standard atomic weight of iridium on three independent determinations of its isotope ratio, which correspond to the following molar mass estimates:

$$M_{\text{obs},1} = 192.216\,77\,(81)\ \text{g/mol}, M_{\text{obs},2} = 192.216\,62\,(29)\ \text{g/mol}, \text{and} \quad M_{\text{obs},3} = 192.217\,63\,(17)\ \text{g/mol}.$$

The standard deviation of these molar mass estimates is 1.6 times greater than the geometric average of the associated standard uncertainties, suggesting that the three measurement results are mutually inconsistent. Cochran's Q test [20], which yields a p-value of 0.0082, confirms the significance of the inconsistency. A statistical measurement model that entertains additional laboratory effects, which are not recognized by the individual measurement uncertainties, can reconcile these three mutually inconsistent results:

$$M_{\text{obs},1} = M + \lambda_1 + \Delta M_1,\ M_{\text{obs},2} = M + \lambda_2 + \Delta M_2,\ \text{and}\ M_{\text{obs},3} = M + \lambda_3 + \Delta M_3,$$

where M denotes the true value of the molar mass, λ_1, λ_2, and λ_3 denote laboratory effects (assumed to be outcomes of random variables with the same Gaussian distribution with mean 0 g/mol and standard deviation τ), and ΔM_1, ΔM_2, and ΔM_3 denote measurement errors, assumed to be independent outcomes of three Gaussian random variables all with the same mean 0 g/mol, and standard deviations equal to the standard uncertainties $u(M_{\text{obs},1})$, $u(M_{\text{obs},2})$, and $u(M_{\text{obs},3})$. These measurement uncertainties are trusted, that is, they are assumed to be based on very large numbers of degrees of freedom.

This statistical measurement model has two parameters of primary interest: M and τ. A naïve evaluation of τ can be obtained similarly to how variance components are often estimated in the analysis of variance [21, Chapter 5]: based on the difference between the squared standard deviation of the measured values of M and the squared geometric average of the standard uncertainties associated with the measured values: $\tau/(\text{g/mol}) = \sqrt{0.000\,54^2 - 0.000\,34^2} = 0.000\,42$. Since this evaluation is based on only three observations, its reliability is questionable, and can be increased either by taking into account additional independent measurements of M, or by putting in play relevant, preexisting knowledge about τ in the form of an informative Bayesian prior distribution for it.

4 Evaluating measurement uncertainty

4.1 Introduction

Type A evaluations involve the application of statistical methods to experimental data, consistent with a measurement model. *Type B* evaluations involve the elicitation of expert knowledge and its encapsulation in

probability distributions that describe states of knowledge about the true values of the inputs. This knowledge can come from authoritative sources such as calibration certificates or technical publications, or from substantive matter experts.

The aforementioned types refer to manners of evaluating uncertainty, not to kinds of measurement error. For example, the uncertainty deriving from the presence of an unknown bias (attributable to a persistent or systematic error), can be evaluated by Type A methods in some cases (for example, via *cross-validation* [22, Ch. 8]), and by Type B methods in other cases (for example, produced by a subject matter expert).

Bottom-up uncertainty evaluations involve the enumeration of all relevant sources of uncertainty, whose contributions are then summarized in an *uncertainty budget*, often supplemented by a *cause-and-effect diagram* [2, Appendix D.3] that describes how they influence the result. *Top-down* uncertainty evaluations involve comparisons between measured values obtained independently of one another, and do not involve the identification and characterization of the sources of the uncertainty, as shown in Examples 3E and 4F.

4.2 Gauss's formula for uncertainty propagation

If the measurement function f in the measurement model $y = f(x_1,\ldots,x_n)$ is differentiable, then the standard uncertainty associated with y can be evaluated approximately using the following generalization (Eq. (13) in [1]) of Gauss's formula [23, §18] for the squared standard uncertainty,

$$u^2(y) \approx \sum_{i=1}^{n} c_i^2 u^2(x_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_i c_j u(x_i) u(x_j) r_{ij}, \tag{1}$$

where $u(x_1),\ldots,u(x_n)$ are the standard uncertainties associated with the estimates of the input quantities, the $\{r_{ij}\}$ are the (Pearson product-moment) correlation coefficients between them, and the *sensitivity coefficients* $c_1,\ldots,c_n$ are the values of the first-order partial derivatives of f with respect to the input quantities, evaluated at $(x_1,\ldots,x_n)$, that is, $c_i = \partial f(x_1,\ldots,x_n)/\partial x_i$, for $i = 1,\ldots,n$ [24]. When the input quantities are uncorrelated, Eq. (1) reduces to

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

Furthermore, when the output quantity, y, is a linear combination of the input quantities, that is, $y = \sum_{i=1}^{n} a_i x_i$, Eq. (1) is exact.

Example 4A: (Water hardness – Gauss's formula) In the case of total water hardness in Example 3C, the sensitivity coefficients are $c_1 = \partial H/\partial\gamma(\mathrm{Ca}^{2+}) = M(\mathrm{CaCO_3})/M(\mathrm{Ca}) \approx 2.5$ and $c_2 = \partial H/\partial\gamma(\mathrm{Mg}^{2+}) = M(\mathrm{CaCO_3})/M(\mathrm{Mg}) \approx 4.1$, and

$$u^2(H) = c_1^2 u^2\left(\gamma\left(\mathrm{Ca}^{2+}\right)\right) + c_2^2 u^2\left(\gamma\left(\mathrm{Mg}^{2+}\right)\right)$$

when we disregard any correlation that may exist between the estimates of $\gamma(\mathrm{Ca}^{2+})$ and $\gamma(\mathrm{Mg}^{2+})$. Such situation arises when calcium and magnesium are measured separately by atomic absorbance spectroscopy. Thus, for water with $\gamma(\mathrm{Ca}^{2+}) = 23.2(8)$ mg/L and $\gamma(\mathrm{Mg}^{2+}) = 9.4(6)$ mg/L, one determined independently of the other, we have

$$u^2(H) = (2.5)^2 \times (0.8\ \mathrm{mg/L})^2 + (4.1)^2 \times (0.6\ \mathrm{mg/L})^2 \approx (3\ \mathrm{mg/L})^2.$$

When the output quantity is a product or a ratio of powers of uncorrelated input quantities, $y = \kappa x_1^{\alpha_1} \ldots x_n^{\alpha_n}$, where κ and $\alpha_1,\ldots,\alpha_n$ are (positive or negative) constants, Gauss's formula takes an easy-to-remember form involving only the relative standard uncertainties: $[u(y)/y]^2 \approx [\alpha_1 u(x_1)/x_1]^2 + \ldots + [\alpha_n u(x_n)/x_n]^2$. In the case of Example 3D, $[u(\rho)/\rho]^2 \approx [u(m_{\mathrm{obs}})/m_{\mathrm{obs}}]^2 + [u(V_{\mathrm{obs}})/V_{\mathrm{obs}}]^2$.

In some applications, the output from the measurement function is a vector with $m > 1$ components, which can be written as a column vector whose elements are the values of m real-valued functions of the n inputs: $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}) \ldots f_m(\boldsymbol{x})]^\top$. Propagating the uncertainty associated with the inputs thus requires the following version of Gauss's formula that produces an approximation to the $m \times m$ covariance matrix of the output vector:

$$\Sigma_y \approx \boldsymbol{J}\Sigma_x\boldsymbol{J}^\top \tag{3}$$

where $\boldsymbol{x} = [x_1 \ldots x_n]^\top$ denotes the n-dimensional (column) vector of the inputs, $\boldsymbol{y}$ denotes the m-dimensional (column) vector of the outputs, and the $m \times n$ Jacobian matrix $\boldsymbol{J}$ has the values of the partial derivatives $\partial f_i(\boldsymbol{x})/\partial x_j$ in row $i = 1,\ldots,m$ and column $j = 1,\ldots,n$. The diagonal elements of Σ_y are the squared standard uncertainties of the outputs, and the off-diagonal elements are covariances between them [25].

Example 4B: (Isotopes of silicon – multivariate uncertainty propagation) Measuring the two isotope ratios of silicon, $R_{29/28}$ and $R_{30/28}$, enables the simultaneous estimation of three isotopic abundances that add up to 1 mol/mol:

$$x_{28} = \frac{1}{1 + R_{29/28} + R_{30/28}}, x_{29} = \frac{R_{29/28}}{1 + R_{29/28} + R_{30/28}}, \text{and } x_{30} = \frac{R_{30/28}}{1 + R_{29/28} + R_{30/28}}.$$

Thus, we have a measurement function of two input quantities whose values are three output quantities, with the following 3×2 Jacobian matrix [26]:

$$\boldsymbol{J} = \begin{bmatrix} \dfrac{\partial x_{28}}{\partial R_{29/28}} & \dfrac{\partial x_{28}}{\partial R_{30/28}} \\ \dfrac{\partial x_{29}}{\partial R_{29/28}} & \dfrac{\partial x_{29}}{\partial R_{30/28}} \\ \dfrac{\partial x_{30}}{\partial R_{29/28}} & \dfrac{\partial x_{30}}{\partial R_{30/28}} \end{bmatrix} = x_{28}^2 \begin{bmatrix} -1 & -1 \\ 1 + R_{30/28} & -R_{29/28} \\ -R_{30/28} & 1 + R_{29/28} \end{bmatrix}.$$

The covariance matrix of the input quantities is

$$\Sigma_{\boldsymbol{R}} = \begin{bmatrix} u^2(R_{29/28}) & ru(R_{29/28})u(R_{30/28}) \\ ru(R_{29/28})u(R_{30/28}) & u^2(R_{30/28}), \end{bmatrix},$$

which includes the correlation coefficient, r, between the estimates of the two input quantities.

Valkiers et al. [27, Table 4] list the following silicon isotope amount ratio measurement results for their Mixture III, $R_{29/28}$ = 0.050 0657(25) mol/mol and $R_{30/28}$ = 0.032 9035(34) mol/mol, but do not provide an estimate of the correlation between them. Assuming that the correlation between them is zero, and applying the formulas above, we obtain x_{28} = 0.923 3873(36) mol/mol, x_{29} = 0.046 2300(22) mol/mol, and x_{30} = 0.030 3827(30) mol/mol. When there are more than two isotope ratios, the uncertainty deriving from lack of knowledge about the correlations between them can be modeled using the LKJ distribution [28].

4.3 Probability distributions

Measurement uncertainty can be represented most thoroughly by a probability distribution. This representation applies equally well to the measurement of quantitative and qualitative properties (Examples 4G, 4L, and 5D illustrate the latter). There are many families of probability distributions, for example, Gaussian (also called *normal*), Laplace, triangular, uniform (also called *rectangular*), gamma, beta, etc. The members of the same family have probability density functions of the same mathematical form, differing only in the values of their parameters. For example, an individual Gaussian distribution is identified by its mean (also known as the expected value) and standard deviation, and individual gamma distributions are identified by the values of their shape and rate parameters.

While the Gaussian distribution gets most of the attention, probabilistic modeling of measurement results often requires choosing other distributions. This choice should be guided by how adequately different distributions encapsulate and convey either a state of knowledge about an input of a measurement model, or an input's sampling variability, as illustrated below and in many examples provided by Possolo and Meija [11].

Counts of various types are modeled using discrete distributions such as the Poisson, negative binomial, or multinomial. This can include numbers of sick patients in a clinical trial, *E. coli* colonies on an agar plate, or scintillations detected by a scintillation counter.

For inputs that can take values in a continuum (for example, a measured amount concentration, whose magnitude can be any non-negative real number), a wide variety of so-called continuous probability distributions are available for modeling. Some of the selection criteria in this case include (i) the range of values that the input can possibly take (for example, whether values can only be positive, or must lie within a bounded interval); (ii) whether values above the center of the distribution are as likely as values below it (symmetry); and (iii) how likely one can expect values that are far away from the center of the distribution (tail heaviness).

It is not always trivial to decide a priori whether a particular input should be modeled using a symmetric or an asymmetric distribution, or how heavy the distribution's tails should be. Some distributions have parameters whose values enable a choice that is responsive to specific features of the measurement data: for example, whether there is skewness (that is, asymmetry) to the left or to the right.

Example 4C: (Modeling asymmetry) Ahrens [29] proposes the lognormal distribution, whose right tail is longer or heavier than the left tail (that is, it exhibits positive skewness), as a model to describe the natural variability of the mass fraction of minor elements in rocks, for example, scandium in samples of Canadian granites. A convolution of Gaussian and exponential distributions (also called exponentially modified Gaussian distribution) is used to model the positive skewness that is commonly displayed by chromatographic peaks. The beta distribution can model negative skewness in the observed purity of high-purity materials. Both the skew normal and the skew *t*-distributions [30] can model either positive or negative skewness.

Example 4D: (Modeling tail heaviness) Both the Laplace distribution and the Student's *t*-distribution can model different degrees of tail heaviness (the smaller the number of degrees of freedom of a Student's *t*-distribution, the heavier its tails are). For example, the probability that a Laplace random variable takes values more than 3 standard deviations away from the mean is 0.014, while the corresponding probability for the Gaussian distributions is only 0.003. For the Student's *t*-distribution with 3 degrees of freedom, the probability of the same event is 0.014. Among discrete distributions, the negative binomial distribution can be used for statistical modeling of counts that are more dispersed than the Poisson distribution allows, as illustrated by Possolo and Meija [11, Pages 41–45].

4.4 Maximum likelihood estimation

For purposes of uncertainty evaluation, we model measurement results as observed values of random variables with specified probability distributions. Hence, we have at our disposal all applicable methods of statistical inference that can produce not only estimates of the quantities of interest but also evaluations of the uncertainty associated with them. The likelihood function (defined below) figures preeminently in many of these methods, both classical and Bayesian.

Suppose the inputs $x_1,\ldots,x_n$ (introduced in 3) can be regarded as outcomes of independent random variables whose probability distributions involve parameters $\theta_1,\ldots,\theta_n$, and whose probability density functions (or probability mass functions) are $p_1,\ldots,p_n$. Letting $\boldsymbol{x}$ denote the set of input quantities, the likelihood function L is

$$L_{\boldsymbol{x}}(\theta_1,\ldots,\theta_n) = p_1(x_1|\theta_1)\ldots p_n(x_n|\theta_n).$$

In practice, the subscript of L indicating the data is often omitted, with the understanding that L is a function of the parameters of the distributions of the input quantities, while the input quantities themselves remain fixed at their measured values. When $p_1,\ldots,p_n$ are probability mass functions, $L_{\boldsymbol{x}}(\theta_1,\ldots,\theta_n)$ is the probability of the data $\boldsymbol{x}$

corresponding to the parameter values $\theta_1,\ldots,\theta_n$. In general, the function L takes its largest values for combinations of values of its arguments, which are the $\{\theta_j\}$, that make the data "most likely."

Example 4E: (SiO_2 in limestone – likelihood function) The certificate of NIST Standard Reference Material 1, Argillaceous Limestone, which was published on July 1st, 1910 [31], lists two determinations of the mass fraction of SiO_2, 0.1811 g/g and 0.1818 g/g, made by W.F. Hillebrand (U.S. Geological Survey) and by C.E. Waters (U.S. National Bureau of Standards), respectively, who employed essentially the same method of wet analysis [32]. If we model these two values as outcomes of Gaussian random variables with the same mean μ and standard deviation σ, then the likelihood function is the product of the values of two Gaussian probability density functions:

$$L(\mu,\sigma) = \frac{\exp\left\{-\frac{(0.1811\ \text{g/g}-\mu)^2}{2\sigma^2}\right\}}{\sqrt{2\pi\sigma^2}} \times \frac{\exp\left\{-\frac{(0.1818\ \text{g/g}-\mu)^2}{2\sigma^2}\right\}}{\sqrt{2\pi\sigma^2}}.$$

This function achieves a maximum when μ equals $\widehat{\mu} = (0.1811\ \text{g/g} + 0.1818\ \text{g/g})/2 = 0.18145\ \text{g/g}$ and σ^2 equals $\widehat{\sigma}^2 = ((0.1811\ \text{g/g} - \widehat{\mu})^2 + (0.1818\ \text{g/g} - \widehat{\mu})^2)/2 = (0.00035\ \text{g/g})^2$.

The maximum likelihood estimate (MLE) of the vector of parameters $\boldsymbol{\theta} = [\theta_1,\ldots,\theta_n]$, denoted $\widehat{\boldsymbol{\theta}}$, is a value of $\boldsymbol{\theta}$ that maximizes the likelihood function, L. (Note that, in some cases, the likelihood function has multiple local maxima, and $\widehat{\boldsymbol{\theta}}$ may not be unique.) For reasons of computational efficiency and stability, it is advantageous to maximize $l = \ln(L)$, which in most applications requires numerical optimization. With great generality, $\widehat{\boldsymbol{\theta}}$ has remarkable properties: the probability distribution of $\widehat{\boldsymbol{\theta}}$ is approximately multivariate Gaussian with mean $\boldsymbol{\theta}$ and covariance matrix $\Sigma_{\widehat{\boldsymbol{\theta}}} \approx -\boldsymbol{H}^{-1}(\widehat{\boldsymbol{\theta}})$, where $\boldsymbol{H}$ (known as the Hessian matrix) is the matrix of second-order partial derivatives of l evaluated at $\widehat{\boldsymbol{\theta}}$, the quality of the approximation increasing with the amount of data that $\widehat{\boldsymbol{\theta}}$ is based on (Theorem 5.1 in [33]).

Example 4F: (Interlaboratory comparison – maximum likelihood estimation) Consider interlaboratory comparison CCQM-K145 with measurements of nickel mass fraction in bovine liver provided by 17 expert laboratories [34]. These results, along with their associated standard measurement uncertainties, are summarized in the following table.

Laboratory	*w*(Ni)/(mg/kg)	Laboratory	*w*(Ni)/(mg/kg)	Laboratory	*w*(Ni)/(mg/kg)
JSI	1.930(110)	INACAL	2.010(60)	PTB	2.077(35)
INMC	1.940(64)	NMIA	2.020(50)	LATU	2.080(59)
GLHK	1.942(92)	NIM	2.022(23)	LGC	2.131(42)
LNE	1.958(75)	NMIJ	2.050(20)	UME	2.150(30)
NIST	1.984(20)	RISE	2.055(52)	HSA	2.180(80)
KRISS	1.993(33)	NRC	2.070(50)		

A laboratory effects model can be used for these data, taking the same general form as in Example 3E, with Gaussian laboratory effects that have mean 0 mg/kg and unknown standard deviation τ. This τ quantifies a source of uncertainty that is often called *dark uncertainty* [35] because its presence is uncovered only once measurement results obtained independently of one another are inter-compared and are found to be more dispersed than their reported, associated uncertainties suggest. The measurement errors are modeled as Gaussian random variables with mean 0 mg/kg and standard deviations equal to the reported uncertainties, which are taken at their face value (that is, they are assumed to be based on large numbers of degrees of freedom). The choice of this statistical model is also supported by the *NIST Decision Tree* [36].

Under the simplifying assumption that the standard uncertainties associated with the measured values are known, the likelihood function has two scalar arguments: the overall mean value of the mass fraction of nickel in the material, μ, and the dark uncertainty, τ. In this case, the likelihood function is a product of 17 Gaussian probability density functions, each pertaining to the result from one participant, similarly to Example 4E. For example, the term corresponding to JSI is

$$\frac{\exp\left\{-\dfrac{(1.930\,\text{mg/kg}-\mu)^2}{2\left(\tau^2+(0.110\,\text{mg/kg})^2\right)}\right\}}{\sqrt{2\pi\left(\tau^2+(0.110\,\text{mg/kg})^2\right)}}.$$

This says that, based on JSI's measurement result alone, the most likely value for μ is the value, 1.930 mg/kg, that JSI measured, and that the uncertainty associated with the measured value is $\sqrt{\tau^2+(0.110\,\text{mg/kg})^2}$, owing to the presence of the dark uncertainty τ.

As noted above, it is preferable to maximize the logarithm of the likelihood function, which, in this case, requires that a numerical optimization method be used. The maximum likelihood estimates are $\widehat{\mu}=2.043\,(15)$ mg/kg and $\widehat{\tau}=0.044\,(14)$ mg/kg. The associated standard uncertainties are based on the Hessian of the log-likelihood function, as described above, even though we have only 17 measurement results. The parametric statistical bootstrap [37], demonstrated in Example 4J, can provide an alternative, generally reliable uncertainty evaluation for the maximum likelihood estimates.

The previous example illustrates how the method of maximum likelihood can be used to combine independent measurements of the same quantity into a consensus value, and how the standard uncertainty associated with this result can be evaluated, in particular when the measurements being combined are mutually inconsistent.

However, the method of maximum likelihood tends to produce biased estimates of the parameters of interest [38, §3], and approximate uncertainty evaluations using the Hessian of the log-likelihood function can seriously underestimate the true variance of the estimates [39, §7], in particular when the number of available measurement results is small (say, less than 15). In general, restricted maximum likelihood estimation [40], for example, as implemented in R package `metafor` [41], or Bayesian procedures like those implemented in the *NIST Decision Tree* [36] and in the *NIST Consensus Builder* [42], are preferable.

Example 4G: (Molnupiravir – maximum likelihood estimation of treatment efficacy) Bernal et al. [43] reported the following final results of the MOVe-OUT clinical trial of the efficacy of molnupiravir against COVID-19, for all the patients who had been randomly allocated to the treatment (800 mg of molnupiravir twice daily for 5 days) and control (placebo) groups: 48 of the $n_1 = 709$ patients in the treatment group and 68 of the $n_0 = 699$ patients in the control group had been hospitalized or died within 29 days of enrolment in the trial.

Even though COVID-19 is an infectious disease, we model these observations as outcomes of binomial distributions whose parameter p (the probability of hospitalization or death) depends on whether the patients were on molnupiravir (p_1) or on placebo (p_0). The corresponding odds of hospitalization or death in the two groups are $O_1 = p_1/(1-p_1)$ and $O_0 = p_0/(1-p_0)$. Since the maximum likelihood estimates of p_1 and p_0 are $\widehat{p}_1 = 48/709 = 0.0677$ and $\widehat{p}_0 = 68/699 = 0.0973$, the odds ratio is $O_1/O_0 = 0.674$. An odds ratio of 1 means that the drug is as effective as the placebo, and the smaller this ratio, the more efficacious the treatment appears to be.

To ascertain whether the odds ratio is significantly smaller than 1, we will evaluate the uncertainty of the log odds ratio, $\ln(O_1/O_0) = -0.395$, which offers the advantage of having a probability distribution that is approximately Gaussian [44]. Since the design of the study guarantees that $\widehat{p}_1$ and $\widehat{p}_0$ are uncorrelated,

$$u^2(\ln(O_1/O_0)) = u^2(\ln O_1) + u^2(\ln O_0).$$

Applying Gauss's formula to each term on the right-hand side of this equation yields

$$u^2(\ln(O_1/O_0)) \approx \frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_0 p_0(1-p_0)} = (0.197)^2.$$

The probability of observing a log odds ratio of −0.395 or smaller (meaning that it speaks at least as favorably of molnupiravir as the trial results do), by chance alone and in the absence of a significant beneficial effect, is approximately equal to the probability of a Gaussian random variable with mean 0 and standard deviation 1 assuming a value smaller than −0.395/0.197 = −2.01, which is only 2.2 %. If the threshold of statistical significance is placed at 5 %, then the results of this trial suggest that molnupiravir is efficacious. However, a subsequent, and a much larger clinical study [45] established that "molnupiravir did not reduce the frequency of COVID-19-associated hospitalizations or death among high-risk vaccinated adults in the community."

4.5 Monte Carlo methods

Monte Carlo simulation methods are often used for uncertainty propagation [46]. They offer several important advantages over the approach described in Section 4.2: (i) do not involve the computation of derivatives, either analytically or numerically; (ii) are applicable even when the measurement models involve markedly non-linear relationships between output and input quantities; and (iii) provide the raw materials necessary to characterize the whole probability distribution of the output quantity.

Monte Carlo methods involve sampling from suitable probability distributions a specified number K of times. The value of K should be sufficiently large to ensure that the resulting uncertainty evaluations (standard uncertainties, coverage intervals, etc.) achieve the required number of significant digits reproducibly.

The Monte Carlo method used to evaluate the uncertainty associated with a measurand in a conventional measurement model, $y = f(x_1,\ldots,x_n)$, involves the following steps:

(a) Specify a joint probability distribution for the input quantities.
(b) Choose a sufficiently large (in the sense specified above) positive integer K and draw a sample of size K from the joint distribution of the input quantities to obtain $(x_{1,1},\ldots,x_{n,1})$, …, $(x_{1,K},\ldots,x_{n,K})$.
(c) Compute $y_1 = f(x_{1,1}, \ldots, x_{n,1}), \ldots, y_K = f(x_{1,K},\ldots, x_{n,K})$, which are a sample from the probability distribution of the output quantity.

The Monte Carlo method produces a large sample $y_1,\ldots,y_K$ drawn from the probability distribution of the output quantity, which fully characterizes the uncertainty associated with y. This sample can be summarized in any one of several different ways, depending on the purpose. For example, for a scalar measurand, a *coverage interval* whose endpoints are the 2.5th and 97.5th sample percentiles of the $y_1,\ldots,y_K$ captures a "typical" value of the output quantity with 95 % probability. In general, such interval is neither a 95 % (classical) confidence interval nor a 95 % (Bayesian) credible interval for the mean of y's distribution. Instead, it is a prediction interval for the value of the output quantity [18, 47].

Example 4H: (Cadmium calibration standard – Monte Carlo uncertainty propagation) Example A1 of the EURACHEM/CITAC Guide CG-4 [2] describes the preparation of a calibration standard for atomic absorption spectroscopy, where a mass m_{Cd} of cadmium with purity w_{Cd} is dissolved in an acid to obtain a solution of volume V. The measurement model expresses the mass concentration of cadmium as $\gamma_{Cd} = w_{Cd}m_{Cd}/V$. The input quantities have the following measured values and standard uncertainties: m_{Cd} = 100.28 mg, $u(m_{Cd})$ = 0.05 mg; w_{Cd} = 0.999 900 g/g, $u(w_{Cd})$ = 0.000 058 g/g; V = 100.00 mL, $u(V)$ = 0.07 mL. Application of Gauss's formula yields $u(\gamma_{Cd})$ = 0.0009 mg/mL.

Now, for the purpose of applying the Monte Carlo method for propagation of distributions, suppose that m_{Cd}, w_{Cd}, and V are modeled as mutually independent random variables with means equal to their measured values and standard deviations equal to their standard uncertainties, and with distributions that are Gaussian, truncated at 0 mg, for the mass (m_{Cd}), beta for the purity (w_{Cd}), and symmetrically triangular for the volume (V).

Given one draw from each of these three distributions, we can compute a value for γ_{Cd} using the measurement model stated above. Doing this 10^7 times produces a Monte Carlo sample of size $K = 10^7$ drawn from the probability distribution of the cadmium concentration. The mean and standard deviation of this sample are γ_{Cd} = 1.0027 mg/mL and $u(\gamma_{Cd})$ = 0.0009 mg/mL. This evaluation of standard uncertainty agrees with the approximate evaluation using Gauss's formula to within one significant digit. However, examination of the Monte Carlo sample also shows that the probability distribution of γ_{Cd} is unimodal and approximately symmetrical, but not Gaussian. Therefore, we evaluate the expanded uncertainty for 95 % coverage that is associated with γ_{Cd} as half of the difference between the 97.5th and 2.5th percentiles of the Monte Carlo sample, obtaining 0.0017 mg/mL.

The parametric statistical bootstrap [37] is a Monte Carlo method suitable to evaluate measurement uncertainty for measurands that are functions of parameters of statistical measurement models. The idea is to estimate parameters of a probability distribution (for example, using least squares, method of moments, or maximum likelihood) and then proceed to sample from this distribution and use the dispersion of the results to characterize the uncertainty of the relevant functions of the parameters.

For example, a linear calibration function is often used to express the relation between values of the mass concentration of an analyte in a solution and instrumental readings – *cf.* Possolo and Meija [11, Page 98]. The slope and intercept, which can be estimated by the method of least squares, are parameters of the probability distribution of the instrumental readings. However, they also determine the value of the measurand that corresponds to any particular value of the instrumental reading. In this case, the parametric bootstrap (called "propagation of distributions" or "Monte Carlo method" in [46]) draws samples from the joint probability distribution of the slope and intercept to evaluate the uncertainty associated with estimates of the measurand. The following Examples 4I and 4J illustrate the use of the parametric bootstrap for uncertainty evaluation.

Example 4I: (Argentometric titration – uncertainty propagation) Consider the measurement of the amount concentration of sodium chloride in aqueous solution by titration with silver nitrate and using potassium chromate as an endpoint indicator (Mohr method), which was introduced in Example 3A, after [48]. The nominal volume of the sodium chloride solution is $V(\mathrm{NaCl})$ = 100 mL, the concentration of the titrant is $c(\mathrm{AgNO_3})$ = 100 mmol/L, of which $V(\mathrm{AgNO_3})$ = 10 mL were consumed until the endpoint was reached. The concentration of the indicator was $c(\mathrm{K_2CrO_4})$ = 1.75 mmol/L. The concentrations of the titrant and of the indicator are both assumed to be known exactly.

Here we will consider only the impact of the uncertainty in the volumetric determinations and in the visual determination of the endpoint. Assuming that $V(\mathrm{NaCl})$ was measured with a Class A volumetric pipette, it can be modeled using a symmetric triangular probability distribution whose probability density function has a symmetrical triangular graph whose base is 0.16 mL wide (the standard deviation is $0.08/\sqrt{6}$ mL). Assuming that $V(\mathrm{AgNO_3})$ was measured using an electronic burette with 0.01 mL resolution, we model $\Delta_V V(\mathrm{AgNO_3})$ using a symmetric triangular distribution whose width is ±0.01 mL, and we model the endpoint detection using a Gaussian distribution whose 95 % coverage interval corresponds to a volume of a single droplet (or standard deviation of 0.025 mL). The measurement model is

$$c(\mathrm{NaCl}) = c(\mathrm{AgNO_3})\frac{V(\mathrm{AgNO_3}) + \Delta_V V(\mathrm{AgNO_3}) + \Delta_E V(\mathrm{AgNO_3})}{V(\mathrm{NaCl}) + \Delta_V V(\mathrm{NaCl})},$$

where $\Delta_V V(\mathrm{AgNO_3})$ and $\Delta_V V(\mathrm{NaCl})$ denote the errors when measuring the volumes of the solutions with the analyte and with the titrant, and $\Delta_E V(\mathrm{AgNO_3})$ denotes the error in the determination of the endpoint. Application of Gauss's formula, or of the *NIST Uncertainty Machine* [49], which implements both Gauss's formula and the parametric bootstrap for the propagation of distributions (Monte Carlo method), yields $c(\mathrm{NaCl})$ = 10.000 mmol/L and $u(c(\mathrm{NaCl}))$ = 0.026 mmol/L, hence a relative standard uncertainty of 0.26 %.

The more refined analysis of the chemical reactions that govern this titration, mentioned in Example 3A, reveals that the difference between the equivalence point and the endpoint is a source of uncertainty whose contribution can be comparable to the uncertainty evaluated above, depending on the concentration of the $\mathrm{K_2CrO_4}$ indicator and the preliminary estimate of the unknown $c(\mathrm{NaCl})$. Instead of modeling this source of uncertainty probabilistically, as above, Meija et al. [14, Eq. (19)] show that the concentration of NaCl can be obtained by solving the following equation for the volume of titrant, $V_E(\mathrm{AgNO_3})$, needed to reach the endpoint:

$$\frac{K_{sp1}}{\sqrt{K_{sp2}\dfrac{V(\mathrm{NaCl}) + V_E(\mathrm{AgNO_3})}{c(\mathrm{K_2CrO_4})V(\mathrm{NaCl})}}} - \sqrt{K_{sp2}\frac{V(\mathrm{NaCl}) + V_E(\mathrm{AgNO_3})}{c(\mathrm{K_2CrO_4})V(\mathrm{NaCl})}} = \frac{c(\mathrm{NaCl})V(\mathrm{NaCl}) - c(\mathrm{AgNO_3})V_E(\mathrm{AgNO_3})}{V(\mathrm{NaCl}) + V_E(\mathrm{AgNO_3})},$$

where K_{sp1} and K_{sp2} are the solubility products of AgCl and $\mathrm{Ag_2CrO_4}$, respectively, whose numerical values are $\log_{10}[K_{sp1}/(\mathrm{mol/L})^2] = -9.75$ and $\log_{10}[K_{sp2}/(\mathrm{mol/L})^3] = -11.9$ [48, Table 1]. The relative difference between the resulting estimate of $c(\mathrm{NaCl})$ and the conventional estimate given above is 0.24 %.

Example 4J: (Michaelis–Menten – Monte Carlo uncertainty propagation) The β-galactosidase enzyme catalyzes the hydrolization of the substrate *o*-nitrophenyl-β-galactoside (ONPG) to produce galactose and *o*-nitrophenol at an initial rate, v_0, that depends on the initial amount concentration of ONPG, c_0, as described by the Michaelis–Menten equation [50, §30A-4] [51, 52]:

$$v_0 = v_{\max}\frac{c_0}{K_M + c_0}.$$

The maximum initial rate, $v_{\max}$, and the Michaelis constant, K_M, can both be estimated from the pairs of values of c_0 and v_0, listed in the following table [53, 54].

c_0 /(mmol L^{-1})	0.05	0.1	0.25	0.5	1	2.5	5	8	20	30
v_0 /(μmol L^{-1} min^{-1})	3	5.2	14.4	30.3	49	86.2	112.6	136.2	170	177.7

The statistical measurement model $v_{0,j} = v_{\max}c_{0,j}/(K_M + c_{0,j}) + \Delta v_{0,j}$, for $j = 1,\ldots,10$, assuming that the errors $\Delta v_{0,1},\ldots,\Delta v_{0,10}$ are like a sample from a Gaussian distribution with mean 0 μmol L^{-1} min^{-1} and standard deviation σ, can be fitted to the data by minimizing $\sum_{j=1}^{n}(v_{0,j} - v_{\max}\,c_{0,j}/(K_M + c_{0,j}))^2$, where $n = 10$ is the number of observations. This nonlinear least squares problem can be solved using a numerical optimizer. The resulting estimates of the parameters in the aforementioned statistical measurement model are $\hat{v}_{\max} = 195$ μmol L^{-1} min^{-1}, and $\hat{K}_M = 3.27$ mmol L^{-1}. The standard deviation of the $\{\Delta v_{0,j}\}$ is estimated as $\hat{\sigma} = 3.09$ μmol L^{-1} min^{-1}, based on 8 degrees of freedom.

To evaluate the uncertainties associated with $\hat{v}_{\max}$ and $\hat{K}_M$, we employ the parametric bootstrap, by repeating the following steps a large number, K, of times, for $k = 1,\ldots,K$:

(1) Randomly select (draw) a sample $e_{1,k},\ldots,e_{n,k}$ from a Student's t-distribution with 8 degrees of freedom, rescaled to have standard deviation $\hat{\sigma}$, and form $v^*_{0,j,k} = \hat{v}_{0,j} + e_{j,k}$, where $\hat{v}_{0,j} = \hat{v}_{\max}c_{0,j}/(\hat{K}_M + c_{0,j})$, ensuring that the probability distribution of the resulting $v^*_{0,j,k}$, for $j = 1,\ldots,10$ is truncated at 0 μmol L^{-1} min^{-1}. We sample from this Student's t-distribution, rather than from a Gaussian distribution, to recognize the number of degrees of freedom that $\hat{\sigma}$ is based on.

(2) Fit the Michaelis–Menten model to the 10 pairs of values of c_0 and v^*_0 using the same procedure mentioned above, to obtain bootstrap estimates of its parameters, $v^*_{\max,k}$ and $K^*_{M,k}$.

The standard deviations of the $\{v^*_{\max,k}\}$ and of the $\{K^*_{M,k}\}$ are the evaluations of $u(v_{\max}) = 3$ μmol L^{-1} min^{-1} and $u(K_M) = 0.18$ mmol L^{-1}. While such uncertainty evaluations can, under favorable circumstances, be obtained using approximations or even exact formulas, the great advantage of the parametric bootstrap is its flexibility that makes it applicable even in challenging situations, regardless of whether appropriate formulas are available.

The parametric bootstrap can also be applied to evaluate the uncertainties associated with the estimates of $v_{\max}$ and of K_M produced by the widely used method proposed by Lineweaver and Burk [55], which is based on the linearized expression $1/v_0 = 1/v_{\max} + (K_M/v_{\max})/c_0$. Even when this relationship is fitted to the data using weights proportional to v_0^4 [53], the resulting estimates are still noticeably biased and their associated uncertainties are larger than their counterparts for the nonlinear model. The choice of weights is highly influential upon the quality of the estimates of the rate parameters, which is yet another reason to abandon the linearized version of the model. In this case, the parametric bootstrap also propagates the uncertainties associated with the weights used in the linear regression that produces the WLB (Weighted Lineweaver–Burk) estimates whose probability densities are depicted in Fig. 1, alongside the probability densities of the estimates (NL) obtained via nonlinear optimization.

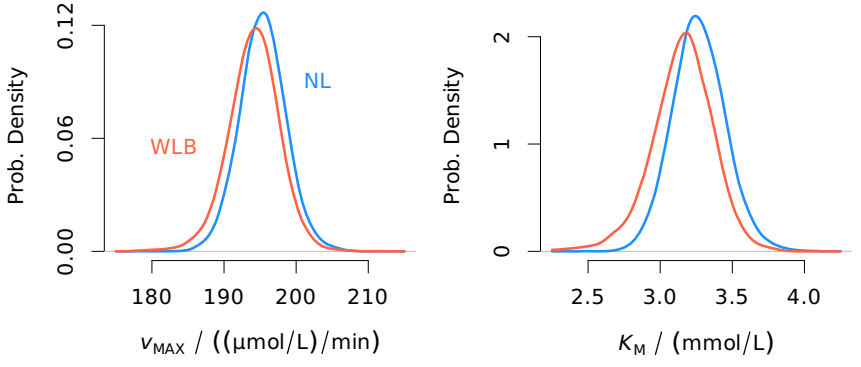

Fig. 1: Probability densities of $v_{\max}$ and K_M estimated by nonlinear regression (NL) and by weighted linear regression (WLB).

4.6 Nonparametric methods

The *nonparametric bootstrap* is even bolder than the parametric bootstrap [37]. It regards the set of available, replicated observations of the quantity of interest, $\{x_1,\ldots,x_n\}$, as if they were an infinitely large sample from an unspecified, underlying distribution, and resample it repeatedly as described below. (Some versions of the bootstrap are applicable when the observations have to be modeled as outcomes of mutually dependent random variables [56].) To evaluate the uncertainty associated with some function, $y = f(x_1,\ldots,x_n)$, of the observations, repeat the following steps for $k = 1,\ldots,K$, a suitably large number, K, of times:

(a) Draw a single value from the set $\{x_1,\ldots,x_n\}$ in such a way that all are equally likely to be drawn, record its value, and return the value drawn to the set. Repeat this n times to obtain a bootstrap sample $\{x^*_{1,k}, \ldots, x^*_{n,k}\}$: that is, each observation x_i is equally likely to be drawn into the kth bootstrap sample, and it may be drawn multiple times into this sample because sampling is done with replacement.

(b) Compute the bootstrap replicate $y^*_k = f(x^*_{1,k}, \ldots, x^*_{n,k})$.

The standard deviation of the bootstrap replicates $y^*_1, \ldots, y^*_K$ is an evaluation of $u(y)$, and the 2.5th and 97.5th percentiles of these replicates are the endpoints of a (generally asymmetrical) 95 % coverage interval for the true value of y.

Nonparametric methods can be useful to produce uncertainty evaluations while making minimal assumptions about the measurement results, for example, to evaluate the uncertainty associated with the median of replicated determinations of the same quantity, which can be done either by application of the nonparametric bootstrap or by inversion of the sign test [57, Chapter 3]. The sign test, and how its inversion produces valid coverage intervals for the median, is detailed in the Supplementary Material.

Example 4K: (Green Seamount basalts – nonparametric bootstrap for the median amount fraction of TiO_2) The following determinations of the mass fraction of TiO_2, expressed in mg/g, in $n = 22$ samples of mid-oceanic ridge basalts from the Green Seamount in the Pacific Ocean near Mexico, described by Allan et al. [58], are listed in the Smithsonian Abyssal Volcanic Glass Data File [59, 60]: 15.2, 14.2, 15.1, 15.0, 12.4, 12.2, 12.7, 14.9, 14.0, 14.0, 14.4, 20.1, 20.5, 11.2, 13.7, 14.1, 12.7, 12.8, 17.9, 11.7, 13.1, 13.5.

It is very unlikely that these determinations are a sample from a Gaussian distribution (the Shapiro–Wilk test [61] of Gaussian shape yields a p-value of 0.003), but they appear to be consistent with a symmetrical distribution (the test of symmetry proposed by Miao et al. [62] yields a p-value of 0.32). Therefore, it may be reasonable to entertain a measurement error model for these determinations of the form $w_{\mathrm{obs},i} = w + \Delta w_i$ for $i = 1,\ldots,n$, similarly to Example 3D, but where w denotes the true mass fraction of TiO_2, and the measurement errors $\Delta w_1,\ldots,\Delta w_n$ are a sample from a Laplace (double exponential) distribution (a test of whether the aforementioned sample could be from a Laplace distribution [63] yields a p-value of 0.99). In these circumstances, the maximum likelihood estimator of w is the median, 14.0 mg/g.

The uncertainty associated with the median can be evaluated using the nonparametric bootstrap, which involves taking a large number, K, of random samples of size $n = 22$, drawn with replacement from the set of 22 observed values above, as explained under 4.6, and computing the median of each such sample.

The reliability of the standard deviation of these bootstrap replicates of the median, as evaluation of the standard uncertainty associated with the median, depends both on the total number of replicates, and on the number of replicates that are numerically different from one another. In general, K should not be small (1000 at least), and the number of numerically different replicates should be large enough so that K is a small fraction of the number, N_B, of possible, different bootstrap samples. In this case, where the determinations comprise $m = 20$ numerically different values, $K = 10^5$ is a tiny fraction of $N_B \approx 7 \times 10^{10}$ [64].

The resulting evaluation of the standard uncertainty of the median mass fraction of TiO_2 was 0.41 mg/g, and a 95 % coverage interval for the same quantity ranged from 12.9 mg/g to 14.7 mg/g. An alternative uncertainty evaluation for the median can be obtained by inversion of the sign test [57, Chapter 3], which yields a 95 % coverage interval ranging from 12.8 mg/g to 14.9 mg/g.

Example 4L: (DNA sequencing – Monte Carlo evaluation of measurement uncertainty for a nominal property) The nucleobase at a particular position in a DNA fragment extracted from NIST Reference Material 8256, Wild-caught Coho Salmon [7], was identified as adenine (A) via high-throughput ("next generation") sequencing. However, the associated probability of erroneous identification, reported in the resulting FASTQ format file, was unusually high: 0.16. In the absence of other information, the probabilities of the four nucleobases at that locus are 0.84 (which is 1 – 0.16) for A (adenine) and 0.16/3 = 0.053 for each of C (cytosine), G (guanine), and T (thymine). This discrete probability distribution fully characterizes the uncertainty in identifying the nucleobase at that position in the DNA. Monte Carlo sampling of similar distributions for the different positions of the DNA fragment generates replicates of the composition of the fragment, which can then be aligned and compared to reference sequences of voucher specimens of closely related species, for purposes of identification of the material (see Example 5D).

4.7 Bayesian methods

Bayesian estimation of model parameters and their associated uncertainties provides the means to blend prior information about the measurand with the fresh information in newly acquired measurement results [65]. In statistical measurement models (refer to Section 3), the measurand is a function of parameters of the probability distribution of the input quantities. The Bayesian approach to estimation and uncertainty evaluation for statistical measurement models involves modeling those parameters as (non-observable) values of random variables, and the measurement data as observed outcomes of random variables whose distributions depend on the unknown values of these parameters [11, Page 204].

Bayesian methods require that the knowledge in hand about those parameters, prior to conducting the measurement experiment, be encapsulated in a probability distribution, which is appropriately called the *prior distribution*. The estimate of the measurand and an evaluation of the associated uncertainty are derived from the conditional distribution of all model parameters given the data. This conditional distribution, called the *posterior distribution*, is obtained by application of Bayes's Rule [66], which states that the posterior probability density function of the parameters is proportional to the product of their joint prior probability density function and the likelihood function.

Example 4M: (Molar mass of iridium – Bayes's rule) Consider the three independent determinations of the molar mass of iridium from Example 3E. The likelihood function is a product of three Gaussian probability density functions, similar to Example 4F. The two parameters of the measurement model, M and τ, are assumed to be independent a priori.

The molar mass M has a Gaussian prior distribution whose mean and standard deviation correspond to the standard atomic weight (relative atomic mass) of iridium, 192.217 ± 0.003, released in 1993 [67]. The prior distribution for τ is half-Cauchy with a median of 0.0004 g/mol, from Example 3E. (The half-Cauchy distribution, which is commonly used for variance components [68], is the Student's t-distribution with 1 degree of freedom, truncated at zero.) This choice of priors amounts to saying that we expect the mean of the three measured values listed in Example 3E to be in line with the 1993 value of the standard atomic weight (relative atomic mass) of iridium, while remaining open to wide discrepancies between the measured values now under consideration.

The probability density function, q, of the joint posterior distribution of M and τ, is

$$q(M,\tau) \propto \underbrace{p_M(M) \times p_\tau(\tau)}_{\text{joint prior}} \times \underbrace{L_{\text{data}}(M,\tau)}_{\text{likelihood}},$$

where the data comprise the three measured values of the molar mass of iridium given in Example 3E ($M_{\text{obs},1}$, $M_{\text{obs},2}$, and $M_{\text{obs},3}$) along with the associated measurement uncertainties $u(M_{\text{obs},1})$, $u(M_{\text{obs},2})$, and $u(M_{\text{obs},3})$. In addition, p_M is the probability density function of the prior distribution of M, which is Gaussian with mean 192.217 g/mol and standard deviation 0.0015 g/mol, and p_τ is the probability density function of the prior distribution of τ, which is the aforementioned half-Cauchy distribution with median 0.000 42 g/mol. (We use the median to specify a particular half-Cauchy distribution because the members of this family of distributions do not have finite mean.)

In practice, Bayes's Rule is used directly only rarely because it involves integrals (or sums) that cannot be evaluated analytically or may be too cumbersome to compute numerically. Instead, Markov Chain Monte Carlo (MCMC) is used to obtain samples from posterior distributions regardless of their complexity [68]. The Random Walk Metropolis algorithm, the Gibbs Sampler, and Hamiltonian Monte Carlo, are widely used versions of MCMC [69, 70].

Example 4N: (Molar mass of iridium – Markov Chain Monte Carlo) We will use MCMC sampling to obtain draws from the posterior distribution of M and τ, as specified in Example 4M. We will employ a particularly simple version of MCMC: the Random Walk Metropolis algorithm with a symmetrical proposal distribution (whose meaning is clarified below). This involves repeating the following steps a large number, K, of times, after making an initial choice of reasonable values, M_0 and τ_0, to assign to M and τ [11], for $k = 1,\ldots,K$:

(1) Compute $p(M_k,\tau_k) = p_M(M_k)p_\tau(\tau_k)L(M_k,\tau_k)$ (suppressing the reference to the data that would usually be made in a subscript for L).
(2) Generate a candidate next pair (M^*, τ^*) in the vicinity of (M_k, τ_k) and calculate $p(M^*, \tau^*)$ as above. We assume that (M^*, τ^*) is a draw from a circularly symmetrical, bivariate Gaussian distribution centered at (M_k, τ_k) with specified, small standard deviations, the so-called *proposal distribution*.
(3) Compute $\alpha = p(M^*, \tau^*)/p(M_k, \tau_k)$.
(4) If $\alpha > 1$, then accept the proposed values of the parameters and set and $\tau_{k+1} = \tau^*$. If $\alpha \leq 1$, then draw a number z from a uniform distribution between 0 and 1 and accept the proposed values only if $\alpha > z$. Otherwise, retain the current values by putting $M_{k+1} = M_k$ and $\tau_{k+1} = \tau_k$.

The initial portion (typically the first half) of the sequence of draws, $(M_1,\tau_1),\ldots,(M_K,\tau_K)$, is usually discarded to ensure that MCMC will have reached equilibrium and is indeed making draws from the intended posterior distribution. The sequence may also be thinned (keeping only every 10th or 25th pair in it, say), before it is used to produce estimates of the parameters and uncertainty evaluations for them.

Applying the version of this algorithm as implemented in R function `MCMCmetrop1R`, which is defined in package `MCMCpack` [71], with $K = 5 \times 10^5$, using the three measurement results for M listed in Example 3E, discarding the first $K/2$ pairs of parameter values, and thinning the rest so as to keep only every 25th, yields an MCMC sample of size 10 000. The mean of the posterior distribution of M is $\widehat{M} = 192.2171\,(4)$ g/mol, and the estimate of dark uncertainty is the median of the posterior distribution of τ, $\widehat{\tau} = 0.0005$ g/mol (we choose the median for τ because the distribution of τ is markedly skewed to the right).

Bayesian models and methods allow us to incorporate the information provided by fresh measurement results along with any relevant preexisting information about measurands or even about other quantities that are relevant for measurement, as illustrated in Examples 4O and 4P. Bayesian techniques also provide flexible ways to honor constraints that some parameters or measurands must satisfy, for example that mass fractions must be positive or that purity cannot exceed 1 g/g, and can even propagate the uncertainty that often surrounds empirical measurement models (for example, concerning the choice of degree for a polynomial that is used as a calibration or analysis function in gas metrology [72]).

Example 4O: (Nitrite in seawater – Bayesian characterization of measurement precision) Four spectrophotometric determinations of the mass fraction of nitrite in a seawater sample were obtained under repeatability conditions of measurement using Griess's method: $w_{\text{obs}}/(\text{mg/kg}) = 0.1514, 0.1523, 0.1531$, and 0.1545.

Based on the past performance of this measurement method, we know that the relative measurement uncertainty is typically 1 % but sometimes could be as low as 0.33 % or as high as 3 %. This prior information about the repeatability standard deviation can be modeled using a gamma distribution centered around 1 % with most (say, 90 %) of its unit of probability spread between 0.33 % and 3 %. Such a gamma distribution has shape $\alpha = 2.619$ and rate $\beta = 1248$ kg/mg.

The statistical measurement model $w_{\text{obs},i} = w + \Delta w_{\text{obs},i}$, for $i = 1,\ldots,4$, relates the observations of the mass fraction of nitrite to their true value, w, where the measurement errors $\{\Delta w_{\text{obs},i}\}$ are assumed to be a sample

from a Gaussian distribution with mean 0 mg/kg and standard deviation σ. The prior information about σ is encapsulated in the aforementioned gamma distribution. A weakly informative Gaussian prior distribution for w, with mean 0 mg/kg and standard deviation 1 mg/kg, truncated at zero, reflects the prior expectation about the order of magnitude of w [11].

MCMC is then used to produce a large sample from the conditional distribution of the parameters, w and σ, given the four observations. The average of the values of w in this sample, 0.1528 mg/kg, is the Bayesian estimate of w, and their standard deviation, 0.0010 mg/kg, is the Bayesian estimate of $u(w)$.

Note that this Bayesian evaluation of the standard uncertainty is about 50 % larger than the conventional Type A evaluation of the standard uncertainty associated with the average of the four replicates. This is not surprising because Bayesian methods usually capture uncertainty contributions more thoroughly than classical methods. This example shows that by taking prior information about an ancillary quantity – measurement precision, σ, in this case – into account can impact significantly the conclusions reached about the measurand of primary interest, which was the mass fraction of nitrites.

Example 4P: (Molnupiravir – skeptical reconsideration via Bayesian modeling) Pogue and Yusuf [73] suggest that "most clinically important interventions are likely to reduce the relative risk of major outcomes, such as myocardial infarction, stroke, or death, by about 10–20 %," and Higgins and Spiegelhalter [74] point out that it is a common occurrence in medical research to see the optimism of early results quenched by subsequent investigations. Example 4G describes an instance of such early optimism about molnupiravir, prompted by the results reported by Bernal et al. [43], which showed a reduction of (0.0973 – 0.0677)/0.0973 = 30 % in the frequency of hospitalizations or death. The subsequent, and much larger, PANORAMIC trial deflated those high hopes [45].

Higgins and Spiegelhalter [74] have shown how a Bayesian approach to the analysis of results from clinical trials can inject a measure of skepticism that counterbalances potentially excessive, early optimism about an intervention. If the treatment and the placebo have approximately the same effect, then the corresponding odds ratio should be 1. Therefore, a reduction of 25 % in the odds of hospitalization or death would put the odds ratio at 0.75. For a Gaussian random variable with mean ln(1) = 0 to take a value smaller than ln(0.75) with 5 % probability, its standard deviation needs to be ln(0.75)/(−1.645) = 0.175, where −1.645 is the 5th percentile a Gaussian distribution with mean zero and unit standard deviation.

Thus, we can use a Gaussian prior distribution with mean 0 and standard deviation 0.175 to convey the prior belief that it is rather unlikely (probability less than 5 %) that a new treatment could reduce the odds of hospitalization or death by more than 25 %. Since the probability distribution of the measured log odds ratio from Example 4G is approximately Gaussian [44], a result from the theory of probability [75, Theorem 1] implies that the posterior distribution of the true value of the log odds ratio also is Gaussian, with mean $(-0.395/0.197^2 + 0/0.175^2)/(1/0.197^2 + 1/0.175^2) = -0.174$ and standard deviation $\sqrt{1/(1/0.197^2+1/0.175^2)} = 0.131$.

The conventional analysis in Example 4G gave molnupiravir's efficacy a z-score of −0.395/0.197 = −2.01, hence suggesting borderline significance for the effect of the treatment. However, once a skeptical prior about the true value of the log-odds ratio is put in play, molnupiravir does not appear to be decisively more efficacious than the placebo (the posterior probability of its being so is only 9.1 %). Therefore, this skeptical, Bayesian reanalysis of the results of the clinical trial reported by Bernal et al. [43] yields similar conclusions to those issuing from the larger, more recent trial reported by Butler et al. [45]: that molnupiravir is not a significantly beneficial treatment for COVID-19 (Fig. 2).

5 Reporting measurement uncertainty

Measurement uncertainties should be reported in a manner that is fit for its intended purpose, which should shape the form of the expression and the number of digits used. Clause 7.2.6 of the GUM [1] recommends that uncertainty evaluations be reported "to at most two significant digits, although in some cases it may be necessary to retain additional digits to avoid round-off errors in subsequent calculations."

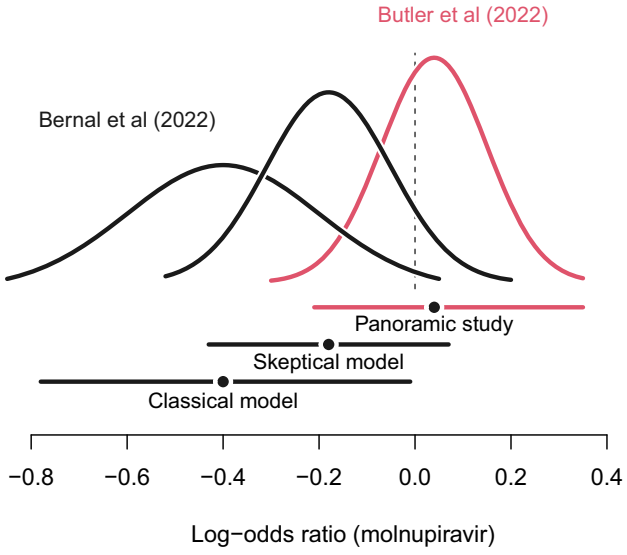

Fig. 2: Evaluating the results of clinical trials with skeptical Bayesian models can lead to more reliable conclusions than classical methods of statistical analysis.

In most cases, specifying a set of values of the measurand believed to include its true value with specified probability suffices as expression of measurement uncertainty. The way such set or interval is built depends on the method used for the uncertainty evaluation as described in Section 4, but necessarily involves some assumption about the probability distribution of the measurand.

If such an interval, $y \pm U_\alpha(y)$, covers the true value of the measurand with a large probability, typically $\alpha > 0.9$, then we say that $U_\alpha(y)$ is an *expanded uncertainty* of y with 100α % coverage probability. By convention, the expanded uncertainty for 95 % coverage is twice the standard uncertainty (see Section 6.5 of [76]). This is motivated by the surprising fact that, in many cases, a coverage interval of the form $y \pm 2u(y)$ does achieve approximately 95 % coverage probability even when the probability distribution is markedly asymmetrical (refer to Pages 68 and A-9 of [77]).

Example 5A: (Measurement result for mass) m = 100.021 47 g with associated standard uncertainty $u(m)$ = 0.000 35 g, also written as m = 100.021 47(35) g. The true value of m lies in (100.021 47 ± 0.000 70) g, or 100.020 77 g ⩽ m ⩽ 100.022 17 g, with approximately 95 % probability.

Example 5B: (Standard atomic weight of boron) The atomic weight (relative atomic mass) of several elements varies significantly among terrestrial materials. Since boron is one of these elements, its standard atomic weight is expressed in the form of an interval, $A_r^o(\mathrm{B}) = [10.806, 10.821]$, signifying that the atomic weight of boron in an unspecified terrestrial material, $A_r(\mathrm{B})$, is $10.806 \leqslant A_r(\mathrm{B}) \leqslant 10.821$, with at least 95 % probability [19].

When the result of an evaluation of measurement uncertainty is intended for use in subsequent uncertainty propagation exercises involving Monte Carlo or Bayesian methods, then the expression of measurement uncertainty should be a fully specified probability distribution for the measurand or a sufficiently large sample drawn from a probability distribution that characterizes the state of knowledge about the measurand.

Example 5C: (Mass fraction of benzoic acid – measurement uncertainty for a quantitative property) The state of knowledge about the mass fraction of benzoic acid in NIST PS1, Primary Standard for Quantitative NMR (Benzoic Acid), is characterized by a beta distribution that approximates the Bayesian posterior distribution with shape parameters α = 888 73.7 and β = 7.11 [78]. The standard deviation of this beta distribution is an evaluation of the standard uncertainty associated with the mass fraction of benzoic acid in NIST PS1, and credible intervals for its true value, of specified probability, can also be derived from the same distribution.

Measurement uncertainty for nominal measurands can be expressed in the form of the Shannon entropy [79, 80] of the corresponding discrete probability distribution.

Example 5D: (Salmon identification – measurement uncertainty for a nominal property) A large number of Monte Carlo replicates (cf. Example 4L) of the composition of a particular DNA fragment extracted from NIST Reference Material 8256, Wild-caught Coho Salmon [7] were aligned and compared to reference sequences from voucher specimens of several closely related species of salmon. The frequencies of the closest matches were as follows: Coho (p_1 = 0.921), Steelhead (p_2 = 0.048), Chinook (p_3 = 0.022), Sockeye (p_4 = 0.005), Pink (p_5 = 0.003), Chum (p_6 = 0.001), and Atlantic (p_7 = 0.000). The Shannon entropy of this discrete probability distribution, $S = -\sum_{i=1}^{7} p_i \ln(p_i) = 0.36$, summarizes the dispersion of the unit of probability over the seven possible identities of the material, hence quantifies the measurement uncertainty for a nominal measurand. The larger the Shannon entropy, the greater the uncertainty. (The largest that the Shannon entropy can be in this case is 1.95, which is achieved when the seven species are deemed to be equally likely. When one is certain about the species in question the Shannon entropy achieves its minimum value, which is 0.) The Shannon entropy is also meaningful for continuous probability distributions. For example, a Gaussian distribution with standard deviation σ has Shannon entropy $S = \ln(\sigma) + \frac{1}{2}\ln(2\pi e)$.

The shortest interval with a specified coverage probability need not be symmetrical around the measured value. The Monte Carlo method can be used to propagate uncertainties expressed asymmetrically, either as asymmetric intervals (Example 5E) or as asymmetric probability distributions, without “symmetrizing” them first [81].

Example 5E: (Asymmetric coverage intervals) The mass fraction of benzoic acid in NIST Primary Standard PS1 for quantitative NMR is $999.92^{+0.04}_{-0.06}$ mg/g with 95 % confidence. This means that the true value of this mass fraction is believed to lie between 999.86 mg/g and 999.96 mg/g with the stated confidence.

Membership of sponsoring bodies

Membership of the IUPAC Analytical Chemistry Division Committee for the period 2020–2021 was as follows: ***President:*** Z. Mester (Canada); ***Secretary:*** D. Craston (UK); ***Vice President:*** D. Shaw (USA); ***Past President:*** J. Labuda (Slovakia); ***Titular Members:*** V. Baranovskaia (Russia); H. Kim (Korea); P. Krystek (Netherlands); C. Magalhães (Portugal); T. Takeuchi (Japan); S. Wiedmer (Finland); ***Associate Members:*** R. Apak (Turkey); J. Barek (Czech Republic); F. Emmerling (Germany); E. Flores (Brazil); I. Kuselman (Israel); H. Li (China/Beijing); ***National Representatives:*** M. F. Camões (Portugal); O. Chailapakul (Thailand); Y.-J. Chen (China/Taipei); A. Felinger (Hungary); B. Hibbert (Australia); S. Ndiaye (Senegal); M. Piston (Uruguay); R. Sha’Ato (Nigeria); L. Torsi (Italy); F. Vanhaecke (Belgium).

Disclaimer

Some commercial entities, equipment, software, or materials may be identified in this contribution to describe an experimental or computational procedure or to illustrate a concept adequately. Such identification does not imply recommendation or endorsement by the IUPAC, the National Research Council (NRC) Canada, or the National Institute of Standards and Technology (NIST), or does it imply that the identified entities, equipment, software, or materials are necessarily the best available for their intended purpose.

Acknowledgments: The authors are grateful to their NIST colleagues Christina Cecelski and Hariharan Iyer for valuable suggestions that stimulated the improvement of an early draft to Debra Ellisor (also from NIST) for sharing information about NIST Reference Material 8256, Wild-caught Coho Salmon and to the anonymous referees for their generous guidance. This technical report is an outcome of IUPAC Project 2015-024-2-500 chaired by D. Brynn Hibbert. The authors are highly appreciative of the support that the IUPAC Analytical Chemistry

Division gave to this project and are very grateful to the late Paul De Bièvre (1933–2016) for initiating the project during the IUPAC General Assembly in Korea (2015) and admiringly dedicate this contribution to his memory.

References

[1] Joint Committee for Guides in Metrology (JCGM). *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement*, International Bureau of Weights and Measures (BIPM), Sèvres, France (2008), URL, https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections.

[2] S. L. R. Ellison, A. Williams (Eds.). *Quantifying Uncertainty in Analytical Measurement*. EURACHEM/CITAC Guide CG-4, QUAM:2012.P1. EURACHEM, 3rd ed. (2012), URL, https://www.eurachem.org.

[3] A. Possolo. *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, National Institute of Standards and Technology, Gaithersburg, MD (2015), NIST Technical Note 1900.

[4] NIST. *NIST Quality Manual for Measurement Services*, National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, Maryland (2019), URL, https://www.nist.gov/nist-quality-system/.

[5] P. Grinberg, K. Nadeau, L. Yang, I. Pihillagawa Gedara, J. Meija, Z. Mester. *AQUA-1: Drinking water Certified Reference Material*, National Research Council Canada, Ottawa, ON (2017).

[6] C. Klein, A. Philpotts. *Earth Materials: Introduction to Mineralogy and Petrology*, Cambridge University Press, Cambridge, UK, 2nd ed. (2017), ISBN 978-1-316-60885-2.

[7] D. L. Ellisor, B. Place, M. Phillips, J. Yen. *Analysis of Seafood Reference Materials: RM 8256, RM 8257, RM 8258, and RM 8259, Wild-Caught Coho Salmon (RM 8256), Aquacultured Coho Salmon (RM 8257), Wild-Caught Shrimp (RM 8258), Aquacultured Shrimp (RM 8259)*, National Institute of Standards and Technology, Gaithersburg, MD (2021), NIST Special Publication 260-214.

[8] J. Michell. *Measurement* **38**, 285 (2005), https://doi.org/10.1016/j.measurement.2005.09.004.

[9] R. White. *Accredit. Qual. Assur.* **16**, 31 (2011), https://doi.org/10.1007/s00769-010-0698-1.

[10] A. Possolo. Measurement. In *Advanced Mathematical and Computational Tools in Metrology and Testing: AMCTM XI, volume 89 of Series on Advances in Mathematics for Applied Sciences*, A. B. Forbes, N.-F. Zhang, A. Chunovkina, S. Eichstädt, F. Pavese, (Eds.), pp. 273–285, World Scientific Publishing Company, Singapore (2018), ISBN 978-981-3274-29-7.

[11] A. Possolo, J. Meija. *Measurement Uncertainty: A Reintroduction*, Sistema Interamericano de Metrologia (SIM), Montevideo, Uruguay, 2nd ed. (2022), ISBN 978-0-660-42866-6. https://doi.org/10.4224/1tqz-b038.

[12] Joint Committee for Guides in Metrology. *Guide to the Expression of Uncertainty in Measurement—Part 6: Developing and Using Measurement Models*, International Bureau of Weights and Measures (BIPM), Sèvres, France (2020), URL, https://www.bipm.org/documents/20126/2071204/JCGM_106_2012_E.pdf/fe9537d2-e7d7-e146-5abb-2649c3450b25. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM GUM-6:2020.

[13] M. F. Camões, G. D. Christian, D. B. Hibbert. *Pure Appl. Chem.* **90**, 563 (2018), https://doi.org/10.1515/pac-2017-0410.

[14] J. Meija, A. M. Michałowska-Kaczmarczyk, T. Michałowski. *Anal. Bioanal. Chem.* **408**, 4469 (2016), https://doi.org/10.1007/s00216-016-9555-3.

[15] E. Pagliano, Z. Mester, J. Meija. *Anal. Chim. Acta* **896**, 63 (2015), https://doi.org/10.1016/j.aca.2015.09.020.

[16] M. S. Oh, H. J. Carroll. *N. Engl. J. Med.* **297**, 814 (1977), https://doi.org/10.1056/NEJM197710132971507.

[17] R. J. Carroll, D. Ruppert, L. A. Stefanski, C. M. Crainiceanu. *Measurement Error in Nonlinear Models—A Modern Perspective*, Chapman & Hall/CRC, Boca Raton, Florida, 2nd ed. (2006).

[18] A. Possolo, H. K. Iyer. *Rev. Sci. Instrum.* **88**, 011301 (2017), https://doi.org/10.1063/1.4974274.

[19] T. Prohaska, J. Irrgeher, J. Benefield, J. K. Böhlke, L. A. Chesson, T. B. Coplen, T. Ding, P. J. H. Dunn, M. Gröning, N. E. Holden, H. A. J. Meijer, H. Moossen, A. Possolo, Y. Takahashi, J. Vogl, T. Walczyk, J. Wang, M. E. Wieser, S. Yoneda, X.-K. Zhu, J. Meija. *Pure Appl. Chem.* **94**, 573 (2022), https://doi.org/10.1515/pac-2019-0603.

[20] W. G. Cochran. *Biometrics* **10**, 101 (1954), https://doi.org/10.2307/3001666.

[21] S. R. Searle, G. Casella, C. E. McCulloch. *Variance Components*, John Wiley & Sons, Hoboken, NJ (2006), ISBN 0-470-00959-4.

[22] F. Mosteller, J. W. Tukey. *Data Analysis and Regression*, Addison-Wesley Publishing Company, Reading, Massachusetts (1977), ISBN 0-201-04854-X.

[23] C. F. Gauss. *Comment. Soc. Regiae Sci. Gottingensis Recentiores, Classis Math.* **5**, 33 (1821). https://resolver.sub.uni-goettingen.de/purl?PPN35283028X_0005_2NS.

[24] H. H. Ku. *J. Res. Natl. Bur. Stand., Sect. C. Eng. Instrum.* **70C**, 263 (1966), https://doi.org/10.6028/jres.070C.025.

[25] Joint Committee for Guides in Metrology. *Evaluation of Measurement Data—Supplement 2 to the "Guide to the Expression of Uncertainty in Measurement"—Extension to any Number of Output Quantities*. International Bureau of Weights and Measures (BIPM), Sèvres, France (2011), URL, https://www.bipm.org/documents/20126/2071204/JCGM_102_2011_E.pdf/6a3281aa-1397-d703-d7a1-a8d58c9bf2a5. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 102:2011.

[26] J. Meija, Z. Mester. *Metrologia* **45**, 53 (2008), https://doi.org/10.1088/0026-1394/45/1/008.

[27] S. Valkiers, G. Mana, K. Fujii, P. Becker. *Metrologia* **48**, S26 (2011), https://doi.org/10.1088/0026-1394/48/2/S04.

[28] D. Lewandowski, D. Kurowicka, H. Joe. *J. Multivariate Anal.* **100**, 1989 (2009), https://doi.org/10.1016/j.jmva.2009.04.008.
[29] L. H. Ahrens. *Geochem. Cosmochim. Acta* **5**, 49 (1954), https://doi.org/10.1016/0016-7037(54)90040-X.
[30] A. Azzalini, A. Capitanio. *The Skew-Normal and Related Families*, Cambridge University Press, Cambridge, UK (2014), ISBN 978-1-107-02927-9.
[31] S. A. Wise, R. L. Watters. *Standard Reference Material 1d, Argillaceous Limestone*, Office of Reference Materials, National Institute of Standards and Technology, Department of Commerce, Gaithersburg, Maryland (2005), URL, www.nist.gov/srm/.
[32] W. F. Hillebrand. *Some Principles and Methods of Rock Analysis, Volume 176 of Bulletin of the United States Geological Survey*, Government Printing Office, Washington, D.C. (1900), URL, https://www.biodiversitylibrary.org/item/310844.
[33] E. L. Lehmann, G. Casella. *Theory of Point Estimation, Springer Texts in Statistics*, Springer-Verlag, New York, NY, 2nd ed. (1998), ISBN 978-1-4419-3130-6.
[34] J. Wang, J. Chao, C. Wei, H. Li, Q. Wang, P. Song, H. Lu, Y. Zhou, Y. Tang, S. Wang, L. Yang, K. Nadeau, I. G. Pihillagawa, M. E. Johnson, L. L. Yu, T. Näykki, T. Sara-Aho, R. P. Zambra, R. Napoli, O. Rienitz, J. Noordmann, C. Pape, J. Towara, C. Tsz-chun, C. Hei-shing, A. Stakheev, V. Dobrovolskiy, T. Stolboushkina, A. Glinkova, S. Taebunpakul, U. Thiengmanee, N. Kaewkhomdee, C. Uribe, E. Carrasco, A. Botha, P. Fisicaro, C. Oster, D. A. Ahumada, J. P. Abella, S. Segura, R. Shin, S. L. P. Deborah, F. Dewi, B. T. M. Kiat, W. Yu Zongrong, L. H. Wah, C. Haraldsson, J. Merrick, L. Antin, I. White, H. Goenaga-Infante, S. Hill, J. Entwisle, R. Jaćimović, T. Zuliani, V. Fajon, Y.-H. Yim, S. W. Heo, K.-S. Lee, J. W. Lee, Y. Lim, T. O. Okumu, M. Ndege, L. Wangui, S. Z. Can, F. Coskun, M. Tunc, P. Giannikopoulou, E. Kakoulides, K. Inagaki, S.-i. Miyashita, H. Klich, R. Jebali, N. Chaaban, L. Bergamaschi, E. Sobina, T. Tabatchikova, P. Migal. *Metrologia* **57**, 08013 (2020), https://doi.org/10.1088/0026-1394/57/1a/08013.
[35] M. Thompson, S. L. R. Ellison. *Accredit. Qual. Assur.* **16**, 483 (2011), https://doi.org/10.1007/s00769-011-0803-0.
[36] A. Possolo, A. Koepke, D. Newton, M. R. Winchester. *J. Res. Natl. Inst. Stand. Technol.* **126**, 126007 (2021), https://doi.org/10.6028/jres.126.007.
[37] B. Efron, R. J. Tibshirani. *An Introduction to the Bootstrap*, Springer-Science+Business Media, Dordrecht, The Netherlands (1993), ISBN 978-0-412-04231-7.
[38] A. L. Rukhin. *J. Res. Natl. Inst. Stand. Technol.* **116**, 539 (2011), https://doi.org/10.6028/jres.116.004.
[39] A. L. Rukhin. *Metrologia* **46**, 323 (2009), https://doi.org/10.1088/0026-1394/46/3/021.
[40] A. Rukhin, B. Biggerstaff, M. Vangel. *J. Stat. Plann. Inference* **83**, 319 (2000), https://doi.org/10.1016/S0378-3758(99)00098-1.
[41] W. Viechtbauer. *J. Stat. Software* **36**, 1 (2010), https://doi.org/10.18637/jss.v036.i03.
[42] A. Koepke, T. Lafarge, A. Possolo, B. Toman. *Metrologia* **54**, S34 (2017), https://doi.org/10.1088/1681-7575/aa6c0e.
[43] A. J. Bernal, M. M. Gomes da Silva, D. B. Musungaie, E. Kovalchuk, A. Gonzalez, V. Delos Reyes, A. Martín-Quirós, Y. Caraco, A. Williams-Diaz, M. L. Brown, J. Du, A. Pedley, C. Assaid, J. Strizki, J. A. Grobler, H. H. Shamsuddin, R. Tipping, H. Wan, A. Paschke, J. R. Butterton, M. G. Johnson, C. De Anda, MOVe-OUT Study Group. *N. Engl. J. Med.* **386**, 509 (2022), https://doi.org/10.1056/NEJMoa2116044.
[44] J. M. Bland, D. G. Altman. *Br. Med. J.* **320**, 1468 (2000), https://doi.org/10.1136/bmj.320.7247.1468.
[45] C. C. Butler, F. D. R. Hobbs, O. A. Gbinigie, N. M. Rahman, G. Hayward, D. B. Richards, J. Dorward, D. M. Lowe, J. F. Standing, J. Breuer, S. Khoo, S. Petrou, K. Hood, J. S. Nguyen-Van-Tam, M. G. Patel, B. R. Saville, J. Marion, E. Ogburn, J. Allen, H. Rutter, N. Francis, N. P. B. Thomas, P. Evans, M. Dobson, T.-A. Madden, J. Holmes, V. Harris, M. E. Png, M. Lown, O. van Hecke, M. A. Detry, C. T. Saunders, M. Fitzgerald, N. S. Berry, L. Mwandigha, U. Galal, S. Mort, B. D. Jani, N. D. Hart, H. Ahmed, D. Butler, M. McKenna, J. Chalk, L. Lavallee, E. Hadley, L. Cureton, M. Benysek, M. Andersson, M. Coates, S. Barrett, C. Bateman, J. C. Davies, I. Raymundo-Wood, A. Ustianowski, A. Carson-Stevens, L.-M. Yu. P. Little, PANORAMIC Trial Collaborative Group. *Lancet* **401**, 281 (2023), https://doi.org/10.1016/S0140-6736(22)02597-1.
[46] Joint Committee for Guides in Metrology. *Evaluation of Measurement Data—Supplement 1 to the "Guide to the Expression of Uncertainty in Measurement"—Propagation of Distributions Using a Monte Carlo Method*, International Bureau of Weights and Measures (BIPM), Sèvres, France (2008), URL, https://www.bipm.org/documents/20126/2071204/JCGM_101_2008_E.pdf/325dcaad-c15a-407c-1105-8b7f322d651c. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 101:2008.
[47] S. Stoudt, A. Pintar, A. Possolo. *J. Res. Natl. Inst. Stand*. **126**, 126004 (2021), https://doi.org/10.6028/jres.126.004.
[48] J. Meija, A. M. Michałowska-Kaczmarczyk, T. Michałowski. *Anal. Bioanal. Chem.* **408**, 1721 (2016), https://doi.org/10.1007/s00216-015-9273-2.
[49] T. Lafarge, A. Possolo. *NCSLI Measure J. Meas. Sci.* **10**, 20 (2015), https://doi.org/10.1080/19315775.2015.11721732.
[50] D. A. Skoog, D. M. West, F. J. Holler, S. R. Crouch. *Fundamentals of Analytical Chemistry*, Brooks/Cole, Belmont, CA, 9th ed. (2014), ISBN 978-0-495-55828-6.
[51] K. A. Johnson, R. S. Goody. *Biochemistry* **50**, 8264 (2011), https://doi.org/10.1021/bi201284u.
[52] L. Michaelis, M. L. Menten. *Biochem. Z.* **49**, 333 (1913).
[53] J. C. Aledo. *Biochem. Mol. Biol. Educ.* **49**, 633 (2021), https://doi.org/10.1002/bmb.21522.
[54] J. C. Aledo. *BMC Bioinf*. **23**, 182 (2022), https://doi.org/10.1186/s12859-022-04729-4.
[55] H. Lineweaver, D. Burk. *J. Am. Chem. Soc.* **56**, 658 (1934), https://doi.org/10.1021/ja01318a036.
[56] S. N. Lahiri. *Resampling Methods for Dependent Data*, Springer-Verlag, New York, NY (2003), ISBN 978-0-387-00928-5.
[57] M. Hollander, D. A. Wolfe, E. Chicken. *Nonparametric Statistical Methods*, John Wiley & Sons, Hoboken, NJ, 3rd ed. (2014), ISBN 978-0-470-38737-5.

[58] J. F. Allan, R. Batiza, P. Lonsdale. Petrology and chemistry of lavas from seamounts flanking the East Pacific Rise axis, 21°N: implications concerning the mantle source composition for both seamount and adjacent EPR lavas. In *Seamounts, Islands, and Atolls, Volume 43 of Geophysical Monograph Series*, B. H. Keating, P. Fryer, R. Batiza, G. W. Boehlert, (Eds.), pp. 255–282, American Geophysical Union (AGU), Washington, D.C. (1987), ISBN 978-0-875-90068-1.
[59] W. G. Melson, T. O'Hearn, E. Jarosewich. *G-cubed* **3**, 1 (2002), https://doi.org/10.1029/2001GC000249.
[60] W. G. Melson, T. O'Hearn, E. Jarosewich. *Smithsonian Abyssal Volcanic Glass Data File*, National Museum of Natural History: Dataset (2020), URL, https://smithsonian.figshare.com/articles/dataset/Smithsonian_Abyssal_Volcanic_Glass_Data_File/11833128.
[61] S. S. Shapiro, M. B. Wilk. *Biometrika* **52**, 591 (1965), https://doi.org/10.2307/2333709.
[62] W. Miao, Y. R. Gel, J. L. Gastwirth. A new test of symmetry about an unknown median. In *Random Walk, Sequential Analysis and Related Topics: A Festschrift in Honor of Yuan-Shih Chow*, A. C. Hsiung, Z. Ying, C.-H. Zhang, (Eds.), pp. 199–214, World Scientific Publishing Company, Singapore (2006).
[63] E. González-Estrada, J. A. Villaseñor. *Stat. Probab. Lett.* **119**, 30 (2016), https://doi.org/10.1016/j.spl.2016.07.003.
[64] A. C. Davison, D. Hinkley. *Bootstrap Methods and their Applications*, Cambridge University Press, Cambridge, UK (1997), ISBN 0-521-57471-4.
[65] A. O'Hagan. The Bayesian approach to statistics. In *Handbook of Probability: Theory and Applications*, T. Rudas (Ed.), Chapter 6, Sage Publications, Thousand Oaks, CA (2008), ISBN 978-1-4129-2714-7.
[66] A. Possolo, B. Toman. *Tutorial for Metrologists on the Probabilistic and Statistical Apparatus Underlying the GUM and Related Documents*, National Institute of Standards and Technology, Gaithersburg, MD (2011), URL, https://www.itl.nist.gov/div898/possolo/TutorialWEBServer/TutorialMetrologists2011Nov09.xht.
[67] J. R. De Laeter, J. K. Böhlke, P. De BiÈVre, H. Hidaka, H. S. Peiser, K. J. R. Rosman, P. D. P. Taylor. *Pure Appl. Chem.* **75**, 683 (2003), https://doi.org/10.1351/pac200375060683.
[68] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin. *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL, 2nd ed. (2003).
[69] C. P. Robert, G. Casella. *Introducing Monte Carlo Methods with R*, Springer, New York, NY (2010), ISBN 978-1-4419-1575-7.
[70] R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, X.-L. Meng (Eds.), Chapter 5, pp. 113–162, Chapman & Hall/CRC, Boca Raton, FL (2011), ISBN 978-1420079418.
[71] A. D. Martin, K. M. Quinn, J. H. Park. *J. Stat. Software* **42**, 22 (2011), https://doi.org/10.18637/jss.v042.i09.
[72] C. E. Cecelski, B. Toman, F.-H. Liu, J. Meija, A. Possolo. *Metrologia* **59**, 045002 (2022), https://doi.org/10.1088/1681-7575/ac711c.
[73] J. Pogue, S. Yusuf. *Lancet* **351**, 47 (1998), https://doi.org/10.1016/S0140-6736(97)08461-4.
[74] J. P. T. Higgins, D. J. Spiegelhalter. *Int. J. Epidemiol.* **31**, 96 (2002), https://doi.org/10.1093/ije/31.1.96.
[75] D. V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint—Part 2, Inference*, Cambridge University Press, Cambridge, UK (1965), ISBN 978-0-521-05563-5.
[76] B. N. Taylor, C. E. Kuyatt. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. NIST Technical Note 1297, National Institute of Standards and Technology, Gaithersburg, MD (1994), URL, https://physics.nist.gov/Pubs/guidelines/TN1297/tn1297s.pdf.
[77] D. A. Freedman. *Statistical Models: Theory and Practice*, Cambridge University Press, New York, NY (2009).
[78] M. A. Nelson, J. F. Waters, B. Toman, B. E. Lang, A. Rück, K. Breitruck, M. Obkircher, A. Windust, K. A. Lippa. *Anal. Chem.* **90**, 10510 (2018), https://doi.org/10.1021/acs.analchem.8b02575.
[79] C. E. Shannon. *Bell System Tech. J.* **27**, 379 (1948).
[80] C. E. Shannon. *Bell System Tech. J.* **27**, 623 (1948).
[81] A. Possolo, C. Merkatas, O. Bodnar. *Metrologia* **56**, 045009 (2019), https://doi.org/10.1088/1681-7575/ab2a8d.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/pac-2022-1203).