NIST Technical Note 1900

Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results

Antonio Possolo

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.TN.1900

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Technical Note 1900

Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results

Antonio Possolo
Statistical Engineering Division
Information Technology Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.TN.1900

October 2015

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Under Secretary of Commerce for Standards and Technology and Director

Orientation

Measurement is an informative assignment of value to quantitative or qualitative properties involving comparison with a standard (§2). Property values that are imperfectly known are modeled as random variables whose probability distributions describe states of knowledge about their true values (§3).

Procedure

(1) **Measurand & Measurement Model.** Define the *measurand* (property intended to be measured, §2), and formulate the *measurement model* (§4) that relates the value of the measurand (output) to the values of inputs (quantitative or qualitative) that determine or influence its value. Measurement models may be:

- *Measurement equations* (§6) that express the measurand as a function of inputs for which estimates and uncertainty evaluations are available (Example E3);
- *Observation equations* (§7) that express the measurand as a function of the parameters of the probability distributions of the inputs (Examples E2 and E14).

(2) **Inputs.** Observe or estimate values for the inputs, and characterize associated uncertainties in ways that are fit for purpose: at a minimum, by standard uncertainties or similar summary characterizations; ideally, by assigning fully specified probability distributions to them, taking correlations between them into account (§5).

(3) **Uncertainty Evaluation.** Select either a *bottom-up* approach starting from an *uncertainty budget* (or, uncertainty analysis), as in TN1297 and in the GUM, or a *top-down* approach, say, involving a proficiency test (§3f). The former typically uses a measurement equation, the latter an observation equation.

(3a) If the measurement model is a measurement equation, and

- The inputs and the output are scalar (that is, real-valued) quantities: use the `NIST Uncertainty Machine` (NUM, `uncertainty.nist.gov`) (§6);
- The inputs are scalar quantities and the output is a vectorial quantity: use the results of the Monte Carlo method produced by the NUM as illustrated in Example E15, and reduce them using suitable statistical analysis software (§6);
- Either the output or some of the inputs are qualitative: use a custom version of the Monte Carlo method (Example E6).

(3b) If the measurement model is an observation equation: use an appropriate statistical method, ideally selected and applied in collaboration with a statistician (§7).

(4) **Measurement Result.** Provide an estimate of the measurand and report an evaluation of the associated uncertainty, comprising one or more of the following (§8):

- *Standard uncertainty* (for scalar measurands), or an analogous summary of the dispersion of values that are attributable to the measurand (for non-scalar measurands);
- *Coverage region:* set of possible values for the measurand that, with specified probability, is believed to include the true value of the measurand;
- *Probability distribution* for the value of the measurand, characterized either analytically (exactly or approximately) or by a suitably large sample drawn from it.

Abbreviations

AIC	Akaike's Information Criterion
ARMA	Auto-regressive, moving average
BIC	Bayesian Information Criterion
DNA	Deoxyribonucleic acid
EIV	Errors-in-variables
GUM	*Guide to the expression of uncertainty in measurement* (Joint Committee for Guides in Metrology, 2008a)
GUM-S1	GUM Supplement 1 (Joint Committee for Guides in Metrology, 2008b)
GUM-S2	GUM Supplement 2 (Joint Committee for Guides in Metrology, 2011)
ICP-MS	Inductively coupled plasma mass spectrometry
ICP-OES	Inductively coupled plasma optical emission spectrometry
ITL	Information Technology Laboratory, NIST
MCMC	Markov Chain Monte Carlo
MQA	Measurement Quality Assurance
NIST	National Institute of Standards and Technology
NUM	NIST Uncertainty Machine (`uncertainty.nist.gov`)
OLS	Ordinary least squares
PSM	Primary standard gas mixture
PCB	Polychlorinated biphenyl
PRT	Platinum resistance thermometer
SED	Statistical Engineering Division (ITL, NIST)
SI	International System of Units (BIPM, 2006)
SRM	NIST Standard Reference Material
TN1297	NIST Technical Note 1297 (Taylor and Kuyatt, 1994)
VIM	*International vocabulary of metrology* (Joint Committee for Guides in Metrology, 2008c)

Typesetting

This *Simple Guide* was typeset using LaTeX as implemented in Christian Schenk's MiKTeX (`www.miktex.org`), using the STIX fonts, a product of the *Scientific and Technical Information Exchange* (STIX) font creation project (`www.stixfonts.org`) of the STI Pub consortium: The American Institute of Physics (AIP), The American Chemical Society (ACS), The American Mathematical Society (AMS), The Institute of Electrical and Electronics Engineering, Inc. (IEEE), and Elsevier.

Purpose & Scope

This document is intended to serve as a succinct guide to evaluating and expressing the uncertainty of NIST measurement results, for NIST scientists, engineers, and technicians who make measurements and use measurement results, and also for our external partners — customers, collaborators, and stakeholders. It supplements but does not replace TN 1297, whose guidance and techniques may continue to be used when they are fit for purpose and there is no compelling reason to question their applicability.

The reader should have some familiarity with the relevant concepts and methods, in particular as described in TN 1297 and in the 1995 version of the GUM. The complete novice should first read the *Beginner's Guide to Uncertainty of Measurement* (Bell, 1999), which is freely available on the World Wide Web.

Since the approximation to standard uncertainty presented as Equation (10) in the GUM was originally introduced and used by Gauss (1823), this *Simple Guide* refers to it, and to the generalized versions thereof that appear as Equations (13) in the GUM and (A-3) in TN 1297, as *Gauss's formula*.

The availability of the `NIST Uncertainty Machine` (NUM) as a service in the World Wide Web (`uncertainty.nist.gov`) (Lafarge and Possolo, 2015) greatly facilitates the application of the conventional formulas for uncertainty propagation, and also the application of the Monte Carlo method that is used for the same purpose. The NUM can reproduce the results of all the examples in TN 1297 and in the GUM.

The scope of this *Simple Guide*, however, is much broader than the scope of both TN 1297 and the GUM, because it attempts to address several of the uncertainty evaluation challenges that have arisen at NIST since the '90s, for example to include molecular biology, greenhouse gases and climate science measurements, and forensic science.

This *Simple Guide* also expands the scope of TN 1297 by recognizing observation equations (that is, statistical models) as measurement models. These models are indispensable to reduce data from key comparisons (Example E10), to combine measurement results for the same measurand obtained by different methods (Example E12), and to characterize the uncertainty of calibration and analysis functions used in the measurement of force (Example E32), temperature (Example E7), or composition of gas mixtures (Examples E17, E18).

Johnson et al. (1994), Johnson et al. (1995), Johnson and Kotz (1972), and Johnson et al. (2005) review all the probability distributions mentioned in this *Simple Guide*, but the *Wikipedia* may be a more convenient, easily accessible reference for them (Wikipedia, 2015) than those authoritative references.

The Examples are an essential complement of the sections in this *Simple Guide*: they are generally arranged in order of increasing complexity of the problem, and of decreasing level of detail that is provided. Complete details, however, are fully documented and illustrated in the R code that is offered separately, as supplementary material. Examples E1–E8 illustrate basic techniques that address many common needs. Metrologists interested in the combination of measurement results obtained either by different methods or laboratories may find Examples E10, E12, E21, E23, and E30 useful.

General Concerns

Some metrologists are concerned with the meaning of probabilistic statements (for example, that specify coverage intervals), and with the related question of whether Bayesian or other statistical methods are best suited for the evaluation of measurement uncertainty.

Bayesian methods should be employed when there is information about the measurand or about the measurement procedure that either originates outside of or predates the measurement experiment, and that should be combined with the information provided by fresh experimental data. A few of the examples in this *Simple Guide* use Bayesian methods (including Examples E19, E10, E25, E22, and E34). The application of Bayesian methods typically is challenging, and often requires collaboration with a statistician or applied mathematician.

O'Hagan (2014) argues persuasively that only a subjective interpretation of probability, reflecting a state of knowledge (either of an individual scientist or of a scientific community), seems capable of addressing all aspects of measurement comprehensively. Since sources of measurement uncertainty attributable to volatile (or, "random") effects cloud states of knowledge about measurands, their contributions can be captured in state-of-knowledge distributions just as well as other contributions to measurement uncertainty.

The subjective interpretation of probability is typically associated with the Bayesian choice that portrays probability as quantification of degrees of belief (Lindley, 2006; Robert, 2007). The term "belief" and derivative terms are used repeatedly in this *Simple Guide*. It is generally understood as "a dispositional psychological state in virtue of which a person will assent to a proposition under certain conditions" (Moser, 1999). Propositional knowledge, reflected in statements like "mercury is a metal", entails belief. Schwitzgebel (2015) discusses the meaning of belief, and Huber and Schmidt-Petri (2009) review degrees of belief.

Questions are often asked about whether it is meaningful to qualify uncertainty evaluations with uncertainties of a higher order, or whether uncertainty evaluations already incorporate all levels of uncertainty. A typical example concerns the average of n observations obtained under conditions of repeatability and modeled as outcomes of independent random variables with the same mean μ and the same standard deviation σ, both unknown *a priori*.

The standard uncertainty that is often associated with such average as estimate of μ equals $s/\sqrt{n}$, where s denotes the standard deviation of the observations. However, it is common knowledge that, especially for small sample sizes, $s/\sqrt{n}$ is a rather unreliable evaluation of $u(\mu)$ because there is considerable uncertainty associated with s as estimate of σ. But then should we not be compelled to consider the uncertainty of that uncertainty evaluation, and so on *ad infinitum*, as if climbing "a long staircase from the near foreground to the misty heights" (Mosteller and Tukey, 1977, Page 2)?

The answer, in this case, with the additional assumption that the observations are like a sample from a Gaussian distribution, is that a (suitably rescaled and shifted) Student's t distribution shortcuts that staircase (Mosteller and Tukey, 1977, 1A) and in fact captures all the shades of uncertainty under consideration, thus fully characterizing the uncertainty associated with the average as estimate of the true mean. Interestingly, this shortcut to that infinite regress is obtained under both frequentist (sampling-theoretic) and Bayesian paradigms for statistical inference.

Questions about the uncertainty of uncertainty pertain to the philosophy of measurement uncertainty, or to epistemology in general (Steup, 2014), and neither to the evaluation nor to the expression of measurement uncertainty. Therefore, they lie outside the scope of this *Simple Guide*.

The following two, more practical questions, also arise often: (a) Are there any better representations of uncertainty than probability distributions? (b) Is there uncertainty associated with a representation of measurement uncertainty? Concerning (a), Lindley (1987) argues forcefully "that probability is the only sensible description of uncertainty and is adequate for all problems involving uncertainty." Aven et al. (2014) discuss differing views. And concerning (b): model uncertainty (Clyde and George, 2004), and ambiguous or incomplete summarization of the dispersion of values of a probability distribution, are potential sources of uncertainty affecting particular representations or expressions of measurement uncertainty.

Nomenclature and Notation

Many models discussed throughout this *Simple Guide* are qualified as being "reasonable". This suggests that most modelers with relevant substantive expertise are likely *a priori* to entertain those models as possibly useful and potentially accurate descriptions of the phenomena of interest, even if, upon close and critical examination, they are subsequently found to be unfit for the purpose they were intended to serve. Similarly, some models are deemed to be "tenable", and are then used, when there is no compelling reason to look for better alternatives: this is often the case only because the data are too scarce to reveal the inadequacy of the models.

And when we say that two models (for example, two probability distributions) are "comparably acceptable", or serve "comparably well" as descriptions of a phenomenon or pattern of variability, we mean that commonly used statistical tests or model selection criteria would fail to find a (statistically) significant difference between their performance, or that any difference that might be found would be substantively inconsequential.

If θ denotes the true value of a scalar quantity that is the object of measurement, for example the temperature of a thermal bath, and we wish to distinguish an estimate of it from its true but unknown value, then we may write $\widehat{\theta} = 23.7\,°\mathrm{C}$, for example, to indicate that 23.7 °C is an estimate, and not necessarily the true value. When it is not important to make this distinction, or when the nature of the value in question is obvious from the context, no diacritical mark is used to distinguish estimate from true value.

However, in all cases we write $u(\theta)$ to denote the associated standard uncertainty because the uncertainty is about the true value of the measurand, not about the specific value that will have been measured for it. To recognize the measurement procedure involved, or generally the context in which the measurement was made, which obviously influence the associated uncertainty, a descriptive subscript may be used. For example $u_{\mathrm{ALEPH,DR}}(m_W)$ $= 0.051\,\mathrm{GeV}/c^2$ denotes the standard uncertainty associated with the mass of the W boson measured by the ALEPH collaboration via direct reconstruction, where c denotes the speed of light in vacuum (The ALEPH Collaboration et al., 2013, Table 7.2) (Example E30). Expanded uncertainties usually are qualified with the corresponding coverage probability as a subscript, as in Example E18, $U_{95\,\%}(c) = 0.40\,\mu\mathrm{mol/mol}$.

Acknowledgments

Adam Pintar, Andrew Rukhin, Blaza Toman, Hari Iyer, John Lu, Jolene Splett, Kevin Coakley, Lo-Hua Yuan, Steve Lund, and Will Guthrie, all from the Statistical Engineering Division (SED), provided guidance, and offered comments and suggestions that led to numerous improvements to earlier versions of this *Simple Guide*.

Wyatt Miller made the observations of surface temperature used in Example E2. Ted Doiron provided guidance and information supporting Example E5. Tom Vetter provided extensive, most insightful comments and suggestions for improvement, and also suggested that an example drawn from Ellison and Williams (2012) would be beneficial (Example E9). Paul Hale reviewed Example E11 and suggested several improvements.

Mike Winchester has stimulated the development and deployment of methods to combine measurement results obtained by different methods, supporting the production of the 3100 series of NIST SRMs. The corresponding technology is illustrated in Example E12 (NIST SRM 3161a): John Molloy shared the data and offered very useful guidance. George Quinn provided the data used in Example E14 and guidance about the underlying study. Harro Meijer (University of Groningen, The Netherlands) stimulated and subsequently reviewed the narrative of Example E16.

Mike Kelley and Jerry Rhoderick provided the data and guidance for Example E18. Example E21 was motivated and influenced by conversations with David Newell, Blaza Toman, and Olha Bodnar (Physikalisch-Technische Bundesanstalt, Berlin, Germany). Larry Lucas and Mike Unterweger kindly reviewed Example E23. Conversations with Graham White (Flinders Medical Center, Australia) stimulated the development of Example E24. Simon Kaplan provided the data used in Example E26 and explanations of the measurement method.

Mark Stiles suggested that an example based on the database mentioned in Example E28 might be useful to draw attention to the fact that *ab initio* calculations are measurements in their own right, hence require qualification in terms of uncertainty evaluations. Mark has also emphasized that consideration of fitness-for-purpose should influence the level of detail required in uncertainty evaluations. Tom Bartel, Kevin Chesnutwood, Samuel Ho, and Ricky Seifarth provided the data used in Example E32, and Tom Bartel also provided extensive guidance supporting this example.

Lesley Chesson (IsoForensics, Salt Lake City, UT) shared uncertainty evaluations for her measurements of δ^2H that are used in Example E33, which reflects joint work with Adriaan van der Veen (Van Swinden Laboratorium (VSL), Delft, The Netherlands), Brynn Hibbert (UNSW, Sydney, Australia), and Juris Meija (National Research Council Canada, Ottawa, Canada).

Juris Meija also suggested the topic of the example in Note 2.6. Bill Luecke suggested that a collection of worked-out examples likely would prove more useful than a very detailed but abstract guide. Chuck Ehrlich offered comments and suggestions for improvement on multiple occasions. Barry Taylor provided very useful guidance to simplify and to increase the accessibility of an early draft of the *Simple Guide*.

Many NIST colleagues examined earlier drafts and provided comments and suggestions for improvement, and many others attended NIST Measurement Services seminars focused on the same topic, and then offered informative assessments in their replies to on-line surveys. Even those comments that expressed complete dissatisfaction or disappointment were very useful and stimulating. Some of the colleagues outside of SED that have kindly provided feedback at some stage or another include: Bob Chirico, Ron Collé, Andrew Dienstfrey, Ted Doiron, David Duewer, Ryan Fitzgerald, Carlos Gonzales, Ken Inn, Raghu Kacker, Paul Kienzle, Bill Luecke, John Sieber, Kartik Srinivasan, Samuel Stavis, Elham Tabassi, Bill Wallace, Donald Windover, and Jin Chu Wu.

Sally Bruce, Jim Olthoff and Bob Watters encouraged the production of this *Simple Guide* as a contribution to the NIST Quality System. David Duewer, Hari Iyer, Mike Lombardi, and Greg Strouse examined it rigorously in the context of its approval for publication: their great knowledge, commitment to rigor, critical thinking, appreciation of the NIST mission, and great generosity, translated into guidance and a wealth of corrections and suggestions for improvement that added great value to the document. Sabrina Springer and Katelynd Bucher provided outstanding editorial guidance and support.

Disclaimers

None of the colleagues mentioned above necessarily underwrites the specific methods used in the examples. Reference to commercial products is made only for purposes of illustration and does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose. Reference to non-commercial products, including the R environment for statistical computing and graphics (R Core Team, 2015) and the NUM, also does not imply that these are the only or necessarily the best tools available for the purposes illustrated in this *Simple Guide*, and only expresses the belief that they are eminently well-suited for the applications where they are employed.

Methods Illustrated in Examples

E1 — WEIGHING. Linear measurement equation, scalar inputs and output, Gauss's formula, Monte Carlo method, analytical evaluation of standard uncertainty.

E2 — SURFACE TEMPERATURE. Observation equation, scalar inputs and output, Gaussian errors, maximum likelihood estimation, nonparametric coverage interval.

E3 — FALLING BALL VISCOMETER. Non-linear measurement equation, scalar inputs and output, Gauss's formula, Monte Carlo method, asymmetric coverage interval and its propagation using gamma approximation.

E4 — PITOT TUBE. Non-linear measurement equation, scalar inputs and output, Gauss's formula, Monte Carlo method, asymmetric coverage interval.

E5 — GAUGE BLOCKS. Linear measurement equation, scalar inputs and output, Gauss's formula, Monte Carlo method, symmetric coverage interval.

E6 — DNA SEQUENCING. Qualitative (categorical) inputs, qualitative and quantitative outputs, quality scores, custom Monte Carlo method, entropy.

E7 — THERMISTOR CALIBRATION. Observation equation for calibration, analysis of variance for model selection, polynomial calibration function and its mathematical inverse (analysis function) for value assignment, roots of cubic equation, Monte Carlo method, simultaneous coverage intervals for analysis function.

E8 — MOLECULAR WEIGHT OF CARBON DIOXIDE. Linear measurement equation, scalar inputs and output, Gauss's formula, Monte Carlo method, analytical characterization of probability distribution of output quantity, trapezoidal distribution of output quantity.

E9 — CADMIUM CALIBRATION STANDARD. Non-linear measurement equation, scalar inputs and output, Gauss's formula, relative uncertainty, Monte Carlo method, asymmetric coverage interval.

E10 — PCB IN SEDIMENT. Observation equation for key comparison, laboratory random effects model, scalar inputs and output, DerSimonian-Laird procedure, parametric statistical bootstrap, Bayesian statistical procedure, Markov Chain Monte Carlo, unilateral degrees of equivalence.

E11 — MICROWAVE STEP ATTENUATOR. Linear measurement equation, scalar inputs and output, beta (arcsine) distribution for input, non-Gaussian distribution of output quantity, Monte Carlo method, bimodal distribution of output quantity, surprising smallest 68 % coverage region.

E12 — TIN STANDARD SOLUTION. Scalar inputs and output, observation equation, random effects model, average, weighted average, DerSimonian-Laird procedure, consensus value, Knapp-Hartung adjustment, Welch-Satterthwaite approximation, parametric statistical bootstrap, lognormal model for the estimate of the between-laboratory variance, uncertainty component for long-term stability.

E13 — THERMAL EXPANSION COEFFICIENT. Non-linear measurement equation, scalar inputs and output, Gaussian or Student's *t* distributions for inputs, Gauss's formula, Monte Carlo method.

E14 — CHARACTERISTIC STRENGTH OF ALUMINA. Observation equation with exponential distributed errors, scalar inputs and output, maximum likelihood estimation of parameters of

Weibull distribution, measurands are either scale parameter of this distribution, or a known function of its scale and shape parameters.

E15 — VOLTAGE REFLECTION COEFFICIENT. Nonlinear measurement equation, complex-valued inputs and output, Monte Carlo method using the NUM, graphical representations of measurement uncertainty.

E16 — OXYGEN ISOTOPES. System of simultaneous observation equations, regression attenuation, errors-in-variables model, Deming regression, non-parametric statistical bootstrap for uncertainty evaluation, statistical comparison of the slopes of two regression lines.

E17 — GAS ANALYSIS. Errors-in-variables regression, calibration and analysis functions, model selection criterion, parametric statistical bootstrap.

E18 — SULFUR DIOXIDE IN NITROGEN. Errors-in-variables regression considering that some Type A evaluations are based on small numbers of degrees of freedom, model selection, analysis function, value assignment of amount fraction of sulfur dioxide to individual cylinders, Type B evaluation of uncertainty component attributable to long-term instability, parametric statistical bootstrap.

E19 — THROMBOLYSIS. Comparison of two medical treatments, observation equation, binary inputs, scalar output (log-odds ratio), approximate uncertainty evaluation for log-odds ratio, elicitation of expert knowledge and its encapsulation in a probability distribution, Bayes's rule, comparison of Bayesian and sampling-theoretic (frequentist) coverage intervals.

E20 — THERMAL BATH. Observation equation with correlated scalar inputs and scalar output, Gaussian auto-regressive moving average (ARMA) time series, model selection, maximum likelihood estimation.

E21 — NEWTONIAN CONSTANT OF GRAVITATION. Observation equation, laboratory random effects model, scalar inputs and output, maximum likelihood estimation with correlated data, parametric statistical bootstrap.

E22 — COPPER IN WHOLEMEAL FLOUR. Observation equation, non-Gaussian data, statistical test for Gaussian shape, robust statistical method, Bayesian approach to robustness.

E23 — TRITIUM HALF-LIFE. Observation equation, scalar inputs and output, meta-analysis, consensus value, random effects model, DerSimonian-Laird procedure, propagating uncertainty associated with between-laboratory differences, Monte Carlo uncertainty evaluation.

E24 — LEUKOCYTES. Observation equation, inputs and outputs are counts (non-negative integers), multinomial model for counts, incorporation of uncertainty component attributable to lack of repeatability either in root sum of squares, or by application of the Monte Carlo method.

E25 — YEAST CELLS. Observation equation, inputs are counts (non-negative integers), scalar output, Poisson distribution for observed counts, Jeffreys's prior distribution, Bayes's rule, analytical posterior probability density, Bayesian coverage interval.

E26 — REFRACTIVE INDEX. Alternative models (non-linear measurement equation, and observation equation) applied to the same data, Edlén's formula, maximum likelihood estimation, Monte Carlo method.

E27 — BALLISTIC LIMIT OF BODY ARMOR. Quantitative and qualitative inputs, scalar output, observation equation, logistic regression for penetration probability, measurement equation for ballistic limit, Monte Carlo method.

E28 — Atomic Ionization Energy. Local-density-functional calculations, observation equation, mixed effects model, coverage intervals for variance components.

E29 — Forensic Glass Fragments. Qualitative (categorical) measurand, observation equation (mixture discriminant analysis), entropy as uncertainty evaluation for categorical measurand, predictive performance of classifier, cross-validation.

E30 — Mass of W Boson. Observation equation, laboratory random effects model, scalar inputs and output, DerSimonian-Laird procedure, Knapp-Hartung correction, parametric statistical bootstrap, lognormal model for uncertainty associated with between-laboratory variance, effective number of degrees of freedom associated with laboratory-specific standard uncertainties.

E31 — Milk Fat. Comparing two different measurement methods graphically, Bland-Altman plot, limits of agreement.

E32 — Load Cell Calibration. Errors-in-variables regression fitted by non-linear, weighted least squares via numerical optimization, calibration and analysis functions, uncertainty evaluation by application of parametric statistical bootstrap.

E33 — Atomic Weight of Hydrogen. Non-linear measurement equation, Monte Carlo method, uniform and Gaussian distributions for inputs, statistical test comparing two measured values.

E34 — Atmospheric Carbon Dioxide. Functional measurand, observation equation, treed Gaussian process, Bayesian procedure to estimate measurand and to evaluate associated uncertainty, simultaneous coverage band for functional measurand.

E35 — Colorado Uranium. Functional measurand, multiple alternative observation equations (local regression, kriging, generalized additive model, multi-resolution Gaussian process model), model uncertainty, model averaging, non-parametric statistical bootstrap.

Sections

1 Grandfathering

(1a) All uncertainty evaluations published as part of measurement results produced in the delivery of NIST measurement services (reference materials, calibrations, and interlaboratory studies that NIST has participated in, including key comparisons) remain valid and need not be redone.

(1b) The conventional procedures for uncertainty evaluation that are described in TN 1297 and in the original version of the GUM may continue to be used going forward when they are fit for purpose and there is no compelling reason to question their applicability.

NOTE 1.1 When the need arises to revise an uncertainty evaluation produced originally according to TN 1297 or to the GUM, suitable alternative procedures illustrated in this *Simple Guide* should be used.

NOTE 1.2 When the results produced by the NUM using Gauss's formula and the Monte Carlo method disagree substantially, then the quality of the approximations underlying Equation (10) in the GUM or Equation (A-3) in TN 1297 is questionable, and the Monte Carlo results should be preferred.

NOTE 1.3 If the function f that appears in the conventional measurement equation is markedly nonlinear in a neighborhood of the estimates of the input quantities that is small by comparison with their standard uncertainties, then Equation (10) in the GUM or Equation (A-3) in TN 1297 may fail to produce a sufficiently accurate approximation to the standard uncertainty of the output quantity, and the Monte Carlo method of the GUM Supplements 1 (GUM-S1) or 2 (GUM-S2) should be used instead.

NOTE 1.4 If the probability distribution of the output quantity demonstrably deviates markedly from Gaussian or Student's t, then the conventional guidance for the selection of coverage factors (GUM Clauses G.2.3 and G.3.2; TN 1297 Subsections 6.2–6.4), may not apply, and coverage intervals or other regions (in particular for multivariate measurands) should be derived from samples drawn from the probability distribution of the measurand by a suitable version of the Monte Carlo method. The NUM computes several of these exact, Monte Carlo intervals for scalar measurands and for the components of vectorial measurands.

2 Measurement is understood in a much wider sense than is contemplated in the current version of the *International vocabulary of metrology* (VIM), and is in general agreement with the definitions suggested by Nicholas and White (2001), White (2011), and Mari and Carbone (2012), to address the evolving needs of measurement science:

> Measurement is an experimental or computational process that, by comparison with a standard, produces an estimate of the true value of a property of a material or virtual object or collection of objects, or of a process, event, or series of events, together with an evaluation of the uncertainty associated with that estimate, and intended for use in support of decision-making.

NOTE 2.1 The property intended to be measured (*measurand*) may be qualitative (for example, the identity of the nucleobase at a particular location of a strand of DNA), or quantitative (for example, the mass concentration of 25-hydroxyvitamin D_3 in NIST SRM 972a, Level 1, whose certified value is 28.8 ng mL^{-1}). The measurand may also be an ordinal property (for example, the Rockwell C hardness of a material), or a function whose values may be quantitative (for example, relating the response of a force transducer to an applied force) or qualitative (for example, the provenance of a glass fragment determined in a forensic investigation).

NOTE 2.2 A measurement standard is a realization or embodiment of the definition of a quantity, including a statement of the value of the quantity and the associated measurement uncertainty (VIM 5.1). This realization may be provided by a measuring system (VIM 3.2), a material measure (VIM 3.6), or a reference material (VIM 5.13). The aforementioned "comparison with a standard" may be direct (for example, using a comparator for the dimensions of gauge blocks), or indirect, via a calibrated instrument (for example, using a force transducer that has been calibrated at NIST).

NOTE 2.3 Measured values are estimates of true values (Ellison and Williams, 2012, F.2.1). An instance of a property has many conceivable values. Whether it has exactly one or more than one true value depends on how the property is defined. The VIM 2.11 defines *true value* of a property as any value of the property that is consistent with the definition of the property (Ehrlich, 2014).

> EXAMPLE: The speed of light in vacuum, and the mass of the sun both have many conceivable values: any positive number of meters per second for the former, and any positive number of kilograms for the latter. The speed of light in vacuum has exactly one true value because one and only one value is consistent with its definition in the SI. The mass of the sun is constantly changing. Even at a particular instant, the mass of the sun depends on how much of its atmosphere is included in its definition.

NOTE 2.4 The VIM 5.1 defines *measurement standard* as a "realization of the definition of a given quantity, with stated value and associated measurement uncertainty, used as a reference", and *reference value* (5.18) as a "value used as a basis for comparison with values of quantities of the same kind."

NOTE 2.5 The evaluation of measurement uncertainty (§3) is an essential part of measurement because it delineates a boundary for the reliability (or trustworthiness) of the assignment of a value (*estimate*) to the measurand, and suggests the extent to which the measurement result conveys the same information for different users in different places and at different times (Mari and Carbone, 2012). For this reason, a measurement result comprises both an estimate of the measurand and an evaluation of the associated uncertainty.

NOTE 2.6 White (2011) explains that the intention to influence an action or to make a decision "is an important reminder that measurements have a purpose that impacts on the definition of the measurand (fitness for purpose), and that a decision carries a risk of being incorrect due to uncertainty in the measurements, and a decision implies a comparison against pre-established performance criteria, a pre-existing measurement scale, and the need for metrological traceability."

> EXAMPLE: The decision-making that measurement supports may arise in any area of science, medicine, economy, policy, or law. The U.S. Code of Federal Regulations (36 C.F.R. §4.23) stipulates that operating or being in actual physical control of a motor vehicle in Federal lands under the administration of the National Park Service is prohibited while the blood alcohol concentration (BAC) is 0.08 grams or more of alcohol per 100 milliliters of blood. A person found guilty of violating this provision will be

punished by a fine or by imprisonment not exceeding 6 months, or both (U.S. 36 C.F.R. §1.3(a)). Gullberg (2012) discusses a case where a person's BAC is measured in duplicate with results of 0.082 g/dL and 0.083 g/dL, but where an uncertainty analysis leads to the conclusion that the probability is only 77 % of the person's BAC actually exceeding the statutory limit.

NOTE 2.7 The term "analyte" is often used in analytical chemistry to identify the substance that is the object of measurement. Since a substance generally has several properties, the measurand is the particular property of the analyte that is intended to be measured: for example, the analyte may be sodium, and the measurand the urine sodium concentration (White and Farrance, 2004).

NOTE 2.8 There may be some unresolvable ambiguity in the definition of the measurand. For example, immunoassays are often used to measure the concentration of vitamin D in serum, using some antibody targeting the relevant forms of the vitamin (cholecalciferol, ergocalciferol, or both). However, the extent of the competition between the vitamin capture antibody and the protein that the vitamin binds to is a source of uncertainty that in some cases may cast some doubt on what a particular immunoassay actually measures (Tai et al., 2010; Farrell et al., 2012). In such cases we speak of *definitional uncertainty* (VIM 2.27), which should be evaluated and propagated as one component of measurement uncertainty whenever it makes a significant contribution to the uncertainty of the result, just like any other uncertainty component.

3 Measurement uncertainty is the doubt about the true value of the measurand that remains after making a measurement. Measurement uncertainty is described fully and quantitatively by a probability distribution on the set of values of the measurand. At a minimum, it may be described summarily and approximately by a quantitative indication of the dispersion (or scatter) of such distribution.

(3a) Measurement uncertainty implies that multiple values of the measurand may be consistent with the knowledge available about its true value, derived from observations made during measurement and possibly also from pre-existing knowledge: the more dispersed those multiple values, the greater the measurement uncertainty (*cf.* VIM 2.26).

(3b) A probability distribution (on the set of possible values of the measurand) provides a complete characterization of measurement uncertainty (Thompson, 2011; O'Hagan, 2014). Since it depicts a state of knowledge, this probability distribution is a subjective construct that expresses how firmly a metrologist believes she knows the measurand's true value, and characterizes how the degree of her belief varies over the set of possible values of the measurand (Ehrlich, 2014, 3.6.1). Typically, different metrologists will claim different measurement uncertainties when measuring the same measurand, possibly even when they obtain the same reading using the same measuring device (because their *a priori* states of knowledge about the true value of the measurand may be different).

(3c) For scalar measurands, measurement uncertainty may be summarized by the standard deviation (*standard uncertainty*) of the corresponding probability distribution, or by similar indications of dispersion (for example, the median absolute deviation from the median). A set of selected quantiles of this distribution provides a more detailed summarization than the standard uncertainty. For vectorial and more general measurands, suitable generalizations of these summaries may be used. For nominal (or, categorical) properties, the entropy of the corresponding probability distribution is one of several possible summary descriptions of measurement uncertainty. None of these summaries, however, characterizes measurement uncertainty completely, each expressing only some particular attributes of the dispersion of the underlying probability distribution of the measurand.

(3d) The plurality of values of the measurand that are consistent with the observations made during measurement may reflect sampling variability or lack of repeatability, and it may also reflect contributions from other sources of uncertainty that may not be expressed in the scatter of the experimental data.

> EXAMPLE: When determining the equilibrium temperature of a thermal bath, repeated readings of a thermometer immersed in the bath typically differ from one another owing to uncontrolled and volatile effects, like convection caused by imperfect insulation, which at times drive the measured temperature above its equilibrium value, and at other times does the opposite. However, imperfect calibration of the thermometer will shift all the readings up or down by some unknown amount.
>
> EXAMPLE: *Dark current* will make the photon flux measured using a charge-coupled device (CCD) appear larger than its true value because the counts generated by the signal are added to the counts generated by dark current. The counts generated by dark current and by the signal both also include counts that represent volatile contributions (Poisson "noise").
>
> The convection effects in the first example, and the Poisson "noise" in the second, are instances of volatile effects. The imperfect calibration of the thermometer in the first example, and the average number of dark current counts that accumulate in each pixel of the CCD per unit of time, are instances of persistent effects, which typically do not manifest themselves in the scatter of readings obtained under conditions of repeatability (VIM 2.20), merely shifting all the readings up or down, yet by unknown amounts.

(3e) In everyday usage, uncertainty and error are different concepts, the former conveying a sense of doubt, the latter suggesting a mistake. Measurement uncertainty and measurement error are similarly different concepts. Measurement uncertainty, as defined above, is a particular kind of uncertainty, hence it is generally consistent with how uncertainty is perceived in everyday usage. But measurement error is not necessarily the consequence of a mistake: instead, it is defined as the difference or distance between a measured value and the corresponding true value (VIM 2.16). When the true value is known (or at least known with negligible uncertainty), measurement error becomes knowable, and can be corrected for.

> EXAMPLE: If 114 V, 212 V, 117 V, 121 V, and 113 V are reported as replicated readings, made in the course of a single day, of the voltage in the same wall outlet in a U.S. residence, then the second value likely is a recording mistake attributable to the transposition of its first two digits, while the dispersion of the others reflects the combined effect of normal fluctuations of the true voltage and of measurement uncertainty.

EXAMPLE: When no photons are allowed to reach a CCD, the true value of the photon flux from any external signal is zero and the bias attributable to dark current is estimated by the counts that accumulate under such conditions.

(3f) *Bottom-up* uncertainty evaluations involve (i) the complete enumeration of all relevant sources of uncertainty, (ii) a description of their interplay and of how they influence the uncertainty of the result, often depicted in a cause-and-effect diagram (Ellison and Williams, 2012, Appendix D)), and (iii) the characterization of the contributions they make to the uncertainty of the result. These elements are often summarized in an uncertainty budget (Note 5.4). *Top-down* uncertainty evaluations, including interlaboratory studies and comparisons with a standard, provide evaluations of measurement uncertainty without requiring or relying on a prior identification and characterization of the contributing sources of uncertainty (Examples E12, E10, E21). Still other modalities may be employed (Wallace, 2010).

NOTE 3.1 Uncertainty is the absence of certainty, and certainty is either a mental state of belief that is incontrovertible for the holder of the belief (like, "I am certain that my eldest son was born in the month of February"), or a logical necessity (like, "I am certain that 426 389 is a prime number"). Being the opposite of an absolute, uncertainty comes by degrees, and measurement uncertainty, which is a kind of uncertainty, is the degree of separation between a state of knowledge achieved by measurement, and the generally unattainable state of complete and perfect knowledge of the object of measurement.

NOTE 3.2 Since measurement is performed to increase knowledge of the measurand, but typically falls short of achieving complete and perfect knowledge of it, measurement uncertainty may be characterized figuratively as the fog of doubt obfuscating the true value of the measurand that measurement fails to lift.

> *In most empirical sciences, the penumbra is at first prominent, and becomes less important and thinner as the accuracy of physical measurement is increased. In mechanics, for example, the penumbra is at first like a thick obscuring veil at the stage where we measure forces only by our muscular sensations, and gradually is attenuated, as the precision of measurements increases.* — Bridgman (1927, Page 36), quoted by Luce (1996)

NOTE 3.3 Bell (1999, Page 1) points out that, to characterize the margin of doubt that remains about the value of a measurand following measurement, we need to answer two questions: *'How big is the margin?' and 'How bad is the doubt?'* In this conformity, and for a scalar measurand for example, it is insufficient to specify just the standard measurement uncertainty without implicitly or explicitly conveying the strength of the belief that the true value of the measurand lies within one or two standard uncertainties of the measured value.

EXAMPLE: The certificate for NIST SRM 972a states explicitly that, with probability 95 %, the mass concentration of 25-hydroxyvitamin D_3 in Level 1 of the material lies within the interval $28.8\,\mathrm{ng\,mL^{-1}} \pm 1.1\,\mathrm{ng\,mL^{-1}}$.

NOTE 3.4 A probability distribution is a mathematical object that may be visualized by analogy with a distribution of mass in a region of space. For example, the Preliminary Reference Earth Model (PREM) (Dziewonski and Anderson, 1981) describes how the earth's mass density varies with the radial distance to the center of the earth. Once this mass density is integrated over the layered spherical shells entertained in PREM that correspond to the main regions in the interior of the earth, we conclude that about 1.3 % of the earth's mass is in the solid inner core, 31 % is in the liquid outer core, 67 % is in the mantle, and 0.7 % is in the crust.

NOTE 3.5 The evaluation of measurement uncertainty is part of the process of measurement quality assurance (MQA). It is NIST policy to maintain and ensure the quality of NIST measurement services (NIST Directive P5400.00, November 20, 2012) by means of a quality management system described in the *NIST Quality Manual* (`www.nist.gov/qualitysystem`). In particular, this policy requires that measured values be accompanied by quantitative statements of associated uncertainties.

NOTE 3.6 According to the NASA Measurement Quality Assurance Handbook, "MQA addresses the need for making correct decisions based on measurement results and offers the means to limit the probability of incorrect decisions to acceptable levels. This probability is termed *measurement decision risk*" (NASA, 2010, Annex 4). Examples of such incorrect decisions include placing a mechanical part in use that is out-of-tolerance, or removing from use a part that, as measured, was found to be out-of-tolerance when in fact it complies with the tolerance requirements. ANSI/NCSL Z540.3 "prescribes requirements for a calibration system to control the accuracy of the measuring and test equipment used to ensure that products and services comply with prescribed requirements" (ANSI/NCSL, 2013).

NOTE 3.7 Calibration (VIM 2.39) is a procedure that establishes a relation between values of a property realized in measurement standards, and indications provided by measuring devices, or property values of artifacts or material specimens, taking into account the measurement uncertainties of the participating standards, devices, artifacts, or specimens. For a measuring device, this relation is usually described by means of a calibration function that maps values of the property realized in the standards, to indications produced by the device being calibrated. However, to use a calibrated device in practice, the (mathematical) inverse of the calibration function is required, which takes an indication produced by the device as input, and produces an estimate of the property of interest as output (Examples E5, E7, E9, E17, E18, and E32).

4 Measurement models describe the relationship between the value of the measurand (*output*) and the values of qualitative or quantitative properties (*inputs*) that determine or influence its value. Measurement models may be measurement equations or observation equations (that is, statistical models).

(4a) A *measurement equation* expresses the measurand as a function of a set of input variables for which estimates and uncertainty evaluations are available.

EXAMPLE: The dynamic viscosity $\mu_{\mathrm{M}} = \mu_{\mathrm{C}}[(\rho_{\mathrm{B}} - \rho_{\mathrm{M}})/(\rho_{\mathrm{B}} - \rho_{\mathrm{C}})](t_{\mathrm{M}}/t_{\mathrm{C}})$ of a solution is expressed as a function of the mass density (ρ_{B}) and travel times (t_{M}, t_{C}) of a ball made to fall through the solution and through a calibration liquid, and of the mass densities of the solution (ρ_{M}) and of the calibration liquid (ρ_{C}) (Example E3).

(4b) An *observation equation* (or, *statistical model*) expresses the measurand as a known function of the parameters of the probability distribution of the inputs.

EXAMPLE: The characteristic strength of alumina is the scale parameter of the Weibull distribution that models the sampling variability of the rupture stress of alumina coupons under flexure (Example E14).

NOTE 4.1 Typically, measurement equations are used in bottom-up uncertainty evaluations, and observation equations are used in top-down uncertainty evaluations (3f).

NOTE 4.2 In general, an observation equation expresses each observed value of an input quantity as a known function of the true value of the measurand and of one or more *nuisance* random variables that represent measurement errors (3e), in such a way that the true value of the measurand appears as a parameter in the probability distribution of the input quantities. (*Cf.* transduction equation in Giordani and Mari (2012, Equation (2)).)

> EXAMPLE: The observation equation in the example under (4b), for the rupture stress s of alumina coupons, may be written explicitly as $\log s = \log \sigma_C + (1/\alpha) \log \varepsilon$, where σ_C (which is the measurand) denotes the *characteristic strength* of the material, and ε denotes measurement error modeled as an exponentially distributed random variable. The mean rupture stress (another possible measurand), is a known function of both parameters, α and σ_C (Example E14).

NOTE 4.3 The following three types of observation equations (or, statistical models) arise often in practice. They may be applicable only to suitably re-expressed data (for example, to the logarithms of the observations, rather than to the observations themselves).

(i) **Additive Measurement Error Model.** Each observation $x = g(y) + \varepsilon$ is the sum of a known function g of the true value y of the measurand and of a random variable ε that represents measurement error (3e). The measurement errors corresponding to different observations may be correlated (Example E20) or uncorrelated (Examples E2 and E14), and they may be Gaussian (Example E2) or not (Examples E22 and E14).

> In some cases, both the measured value and the measurement error are known to be positive, and the typical size (but not the exact value) of the measurement error is known to be proportional to the true value of the measurand. The additive measurement error model may then apply to the logarithms of the measured values.

(ii) **Random Effects Model.** The value $x_i = y + \lambda_i + \varepsilon_i$ measured by laboratory i, or using measurement method i, is equal to the true value y of the measurand, plus the value λ_i of a random variable representing a laboratory or method effect, plus the value ε_i of a random variable representing measurement error, for $i = 1, \dots, m$ laboratories or methods. This generic model has many specific variants, and can be fitted to data in any one of many different ways (Brockwell and Gordon, 2001; Iyer et al., 2004). This model should be used when combining measurement results obtained by different laboratories, including interlaboratory studies and key comparisons, (Examples E10 and E21) or by different measurement methods (Example E12), because it recognizes and evaluates explicitly the component of uncertainty that is attributable to differences between laboratories or methods, the so-called *dark uncertainty* (Thompson and Ellison, 2011).

(iii) **Regression Model.** The measurand y is a function relating corresponding values of two quantities at least one of which is corrupted by measurement error (Examples E7, E17, E18, E32, and E34), for example when y is a third-degree polynomial and the amount-of-substance fraction of a gaseous species in a mixture is given by $x = y(r) + \varepsilon$, where r denotes an instrumental indication and the random variable ε denotes measurement error. Many calibrations involve the determination of such function y using methods of statistical regression analysis (Examples E7, E32).

NOTE 4.4 In many cases, several alternative statistical models may reasonably be entertained that relate the observations to the true value of the measurand. Even when a criterion is used to select the "best" model, the fact remains that there is *model uncertainty*, which should be characterized, evaluated, and propagated to the uncertainty associated with the estimate of the measurand, typically using Monte Carlo or Bayesian methods. At a minimum, the sensitivity of the results to model choice should be evaluated (7c).

> EXAMPLE: Examples E17 and E18 illustrate the construction of a gas analysis function that takes as input an instrumental indication, and produces as output an estimate of the amount-of-substance fraction of an analyte: in many applications, this function is often assumed to be a polynomial but there is uncertainty about its degree, which is a form of model uncertainty.

NOTE 4.5 The probability distribution that is used to describe the variability of the experimental data generally is but one of several, comparably acceptable alternatives that could be entertained for the data: this plurality is a manifestation of model uncertainty (Clyde and George, 2004).

> EXAMPLE: In Example E14, the natural variability of the rupture stress of alumina coupons may be described comparably well by lognormal or by Weibull probability distributions. And in Example E2 an extreme value distribution may be as tenable a model as the Gaussian distribution that is entertained there.

5 Uncertainty evaluations for inputs to measurement models are often classified into Type A or Type B depending on how they are performed:

- ***Type A*** evaluations involve the application of statistical methods to experimental data, consistently with a measurement model;
- ***Type B*** evaluations involve the elicitation of expert knowledge (from a single expert or from a group of experts, also from authoritative sources including calibration certificates, certified reference materials, and technical publications) and its distillation into probability distributions (or summaries thereof that are fit for purpose) that describe states of knowledge about the true values of the inputs.

(5a) The GUM defines Type B evaluations more broadly than above, to comprise any that are not derived from "repeated observations". In particular, even if the pool of information that the evaluation draws from consists of "previous measurement data", the GUM still classifies it as of Type B, apparently weighing more heavily the "previous" than the "data". Even though the definition above does not specify what the expert knowledge may have been derived from, by insisting on "elicitation" it suggests that the source is (subjective) knowledge. When this knowledge is drawn from a group of experts, the resulting probability distribution will have to capture not only the vagueness of each expert's knowledge, but also the diversity of opinions expressed by the experts (Baddeley et al., 2004; Curtis and Wood, 2004; O'Hagan et al., 2006).

(5b) *The purpose of the Type A and Type B classification is to indicate the two different ways of evaluating uncertainty components and is for convenience of discussion* *only; the classification is not meant to indicate that there is any difference in the nature of the components resulting from the two types of evaluation. Both types of evaluation are based on probability distributions* — GUM 3.3.4.

Unfortunately, this purpose is often ignored, and the classification into types is often erroneously interpreted as suggesting one of more of the following: (i) Type A evaluations and Type B evaluations are not comparably reliable; (ii) Type A evaluations are for uncertainty components attributable to "random" effects; (iii) Type B evaluations are for uncertainty components attributable to "systematic" effects.

Therefore, rather than using a classification that is much too often misunderstood or misapplied, we recommend that the original source of the uncertainty evaluation be stated explicitly, and described with a level of detail fit for the purpose that the evaluation is intended to serve: experimental data (even if more than one step removed from the immediate source of the uncertainty evaluation), meta-analysis (Cooper et al., 2009), literature survey, expert opinion, or mere guess.

> EXAMPLE: When the user of NIST SRM 1d (Wise and Watters, 2005b) extracts from the corresponding certificate the expanded uncertainty, 0.16 %, associated with the mass fraction, 52.85 %, of CaO in the material (argillaceous limestone), according to the GUM this expanded uncertainty becomes the result of a Type B evaluation for the user of the certificate even though it rests entirely on a statistical analysis of experimental data obtained by multiple laboratories using different analytical methods.

(5c) Irrespective of their provenance and of how they are evaluated, uncertainty components should all be treated alike and combined on an equal footing, which is how TN1297, the GUM, and the NUM treat them. Characterizing them via fully specified probability distributions (or via samples from these distributions) facilitates such uniform treatment, in particular when both quantitative and qualitative inputs together determine the value of the measurand.

(5d) Classifying the methods used to evaluate uncertainty according to how they operate is certainly easier and less controversial than classifying the sources or components of uncertainty according to their nature. For example, declaring that an uncertainty component is either *random* or *systematic* involves a judgment about its essence and presupposes that there is a widely accepted, common understanding of the meaning of these terms. Both the GUM and TN1297 appropriately eschew the use of these qualifiers, and this *Simple Guide* reaffirms the undesirability of their use.

> *The nature of an uncertainty component is conditioned by the use made of the corresponding quantity, that is, on how that quantity appears in the mathematical model that describes the measurement process. When the corresponding quantity is used in a different way, a "random" component may become a "systematic" component and vice versa. Thus the terms "random uncertainty" and "systematic uncertainty" can be misleading when generally applied*
> — TN1297, Subsection 2.3 (Pages 1–2).

(5e) For purposes of uncertainty evaluation, in particular considering the flexibility afforded by Monte Carlo methods, it is preferable to classify uncertainty components according to the behavior of their effects, as either *persistent* or *volatile*.

> EXAMPLE: When calibrating a force transducer, the orientation of the transducer relative to the loading platens of the deadweight machine is a persistent effect because a change in such orientation may shift the transducer's response up or down at all set-points of the applied force, by unknown and possibly variable amounts, but all in the same direction (Bartel, 2005, Figure 5).

(5f) Uncertainty evaluations should produce, at a minimum, estimates and standard uncertainties of all the inputs when these are scalar quantities, or suitable proxies of the standard uncertainties for other measurands. Ideally, however, these evaluations should produce fully specified probability distributions (or samples from such distributions) for the inputs.

(5g) Both types of measurement models (measurement equations and observation equations) involve input variables whose values must be estimated and whose associated uncertainties must be characterized.

> EXAMPLE: In Example E11 the uncertainty associated with the output is evaluated using a bottom-up approach. The measurement model is a measurement equation. Some of its inputs are outputs of measurement models used previously, and the associated uncertainties were evaluated using Type A methods. Other inputs had their uncertainties evaluated by Type B methods.
>
> EXAMPLE: In Example E10 the measurement model is an observation equation, and the uncertainty associated with the output is evaluated using a top-down approach. The inputs are measured values, associated uncertainties, and the numbers of degrees of freedom that these uncertainties are based on.

(5h) In the absence of compelling reason to do otherwise, (univariate or multivariate) Gaussian probability distributions may be assigned to quantitative inputs (but refer to (5i) next for an important exception). Under this modeling choice, and in many cases, it is likely that Gauss's formula and the Monte Carlo method will lead to similar evaluations of standard uncertainty for scalar measurands specified by measurement equations. Discrete uniform distributions (which assign the same probability to all possible values) may be appropriate for qualitative inputs, but typically other choices will be preferable.

(5i) If the measurement model is a measurement equation involving a ratio, it is inadvisable to assign a Gaussian distribution to any variable that appears in the denominator because this induces an infinite variance for the ratio. If the variable is positive and its coefficient of variation (ratio of standard uncertainty to mean value of the variable) is small, say, no larger than 5 %, then a lognormal distribution with the same mean and standard deviation is a convenient alternative that avoids the problem of infinite variance.

(5j) Automatic methods for assignment of distributions to inputs (for example, "rules" based on maximum entropy considerations) should be avoided. In all cases, the choice should be the result of deliberate model selection exercises, informed by specific knowledge about the inputs, and taking into account the pattern of dispersion apparent in relevant experimental data. The advice that the GUM (3.4.8) offers is particularly relevant here: that any framework for assessing uncertainty *cannot substitute for critical thinking, intellectual honesty and professional skill. The evaluation of uncertainty is neither a routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand and of the measurement.*

(i) If the range of the values that a scalar input quantity may possibly take is bounded, then a suitably rescaled and shifted beta distribution (which includes the rectangular distribution as a special case), a triangular distribution, or a trapezoidal distribution may be a suitable model for the input quantity.

(ii) First principles considerations from the substantive area of application may suggest non-Gaussian distributions (Example E11).

(iii) Student's *t*, Laplace, and hyperbolic distributions are suitable candidates for situations where large deviations from the center of the distribution are more likely than under a Gaussian model.

(iv) Lognormal, gamma, Pareto, Weibull, and generalized extreme value distributions are candidate models for scalar quantities known to be positive and such that values larger than the median are more likely than values smaller than the median.

(v) Distributions for vectorial (multivariate) quantities may be assembled from models for the distributions of the components of the vector, together with the correlations between them, using copulas (Possolo, 2010), even though doing so requires making assumptions whose adequacy may be difficult to judge in practice.

(vi) Several discrete distributions may be useful to express states of knowledge about qualitative inputs. In many cases, the probability distribution that best describes the metrologist's state of knowledge about the true value of a qualitative property will not belong to any particular family of distributions (Example E6). In some cases a uniform discrete distribution (that is, a distribution that assigns the same probability to each of a finite set of values) is appropriate, Example E29). The binomial (Example E27), multinomial (Example E24), negative binomial, and Poisson (Example E25) distributions are commonly used.

(vii) Since Type A evaluations typically involve fitting a probability model to data obtained under conditions of repeatability (VIM 2.20), the selection of a probability model, which may be a mixture of simpler models (Benaglia et al., 2009), should follow standard best practices for model selection (Burnham and Anderson, 2002) and for the verification of model adequacy, ideally applied in collaboration with a statistician or applied mathematician.

NOTE 5.1 Possolo and Elster (2014) explain how to perform Type A and Type B evaluations, and illustrate them with examples.

NOTE 5.2 O'Hagan et al. (2006) provide detailed guidance about how to elicit expert knowledge and distill it into probability distributions. The European Food Safety Authority has endorsed these methods of elicitation for use in uncertainty quantification associated with dietary exposure to pesticide residues (European Food Safety Authority, 2012). O'Hagan (2014) discusses elicitation for metrological applications. Baddeley et al. (2004) provide examples of elicitation in the earth sciences. Example E19 describes an instance of elicitation.

NOTE 5.3 In many cases, informal elicitations suffice: for example, when the metrologist believes that a symmetrical triangular distribution with a particular mean and range describes his state of knowledge about an input quantity sufficiently accurately for the intended purpose (say, because the number of significant digits required for the results do not warrant the effort of eliciting a shifted and scaled beta distribution instead). The *Sheffield Elicitation Framework* (SHELF) (O'Hagan, 2012), and the *MATCH Uncertainty Elicitation Tool* (Morris et al., 2014) facilitate a structured approach to elicitation: `optics.eee.nottingham.ac.uk/match/uncertainty.php`.

NOTE 5.4 The results of uncertainty evaluations for inputs that are used in a bottom-up evaluation of measurement uncertainty, should be summarized by listing the identified sources of uncertainty believed to contribute significantly to the uncertainty associated with the measurand, and by characterizing each uncertainty contribution. (This summary is commonly called an *uncertainty budget*, or an uncertainty analysis.) At a minimum, such characterization involves specifying the standard uncertainty for scalar measurands, or its analog for measurands of other types. Ideally, however, a probability distribution should be specified that fully describes the contribution that the source makes to the measurement uncertainty.

EXAMPLE: Sources of uncertainty (uncertainty budget, or uncertainty analysis) for the gravimetric determination of the mass fraction of mercury in NIST SRM 1641e (Mercury in Water), adapted from Butler and Molloy (2014, Table 1), where "DF" denotes the number of degrees of freedom that the standard uncertainty is based on, and "MODEL" specifies the probability model suggested for use in the NUM. The Student t distribution has 12.5 degrees of freedom, computed using the Welch-Satterthwaite formula as described in TN1297 (B.3) and in the GUM (G.4), and it is shifted and rescaled to have mean and standard deviation equal to the estimate and standard uncertainty listed for w_{3133}. The measurement equation is $w_{1641e} = m_{3133} w_{3133} m_{\text{spike}} / (m_{\text{spiking soln}} m_{1641e})$.

INPUT	ESTIMATE	STD. UNC.	DF	MODEL
m_{3133}	1.0024 g	0.0008 g	∞	Gaussian
w_{3133}	9.954×10^6 ng/g	0.024×10^6 ng/g	12.5	Student t
$m_{\text{spiking soln}}$	51.0541 g	0.0008 g	∞	Gaussian
m_{spike}	26.0290 g	0.0008 g	∞	Gaussian
m_{1641e}	50 050.6 g	0.1 g	∞	Gaussian

6 Uncertainty evaluation for measurands defined by measurement equations

(6a) If the inputs are quantitative and the output is a scalar quantity, then use the NUM (Lafarge and Possolo, 2015), available on the World Wide Web at `uncertainty.nist.gov`, with user's manual at `uncertainty.nist.gov/NISTUncertaintyMachine-UserManual.pdf`.

NOTE 6.1 The NUM implements both the Monte Carlo method described in the GUM-S1, and the conventional Gauss's formula for uncertainty propagation, Equations (A-3) in TN1297 and (13) in the GUM, possibly including correlations, which the NUM applies using a copula (Possolo, 2010).

NOTE 6.2 The NUM requires that probability distributions be assigned to all input quantities.

(6b) When the Monte Carlo method and the conventional Gauss's formula for uncertainty propagation produce results that are significantly different (judged considering the purpose of the uncertainty evaluation), then the results from the Monte Carlo method are preferred.

(6c) If the inputs are quantitative and the output is a vectorial quantity, then use the Monte Carlo method, as illustrated in Example E15.

(6d) If some of the inputs are qualitative, or the output is neither a scalar nor a vectorial quantity, then employ a custom Monte Carlo method, ideally selected and applied in collaboration with a statistician or applied mathematician.

> EXAMPLES: Examples E6, E27, and E29 illustrate how uncertainty may be propagated when some of the inputs or the output are qualitative, and Examples E17, E18, E32 and E34 do likewise for a functional measurand.

7 Uncertainty evaluation for measurands defined by observation equations starts from the realization that observation equations are statistical models where the measurand appears either as a parameter of a probability distribution, or as a known function of parameters of a probability distribution. These parameters need to be estimated from experimental data, possibly together with other relevant information, and the uncertainty evaluation typically is a by-product of the statistical exercise of fitting the model to the data.

> EXAMPLE: In Example E14, one measurand is the characteristic strength of alumina, which appears as the scale parameter of a Weibull distribution. This parameter is estimated by the method of maximum likelihood, which produces not only an estimate of this scale parameter, but also an approximate evaluation of the associated uncertainty. Another measurand is the mean rupture stress, which is a known function of both the scale and shape parameters of that Weibull distribution.

(7a) Observation equations are typically called for when multiple observations of the value of the same property are made under conditions of repeatability (VIM 2.20), or when multiple measurements are made of the same measurand (for example, in an interlaboratory study), and the goal is to combine those observations or these measurement results.

> EXAMPLES: Examples E2, E20, and E14 involve multiple observations made under conditions of repeatability. In Examples E12, E10, and E21, the same measurand has been measured by different laboratories or by different methods.

(7b) In all cases, the adequacy of the model to the data must be validated. For example, when fitting a regression model (Note 4.3) we should examine plots of residuals (differences between observed and fitted values) against fitted values to determine whether any residual structure is apparent that the model failed to capture (Fox and Weisberg, 2011, Chapter 6). QQ-plots (Wilk and Gnanadesikan, 1968) of the residuals should also be examined, to detect possibly significant inconsistencies with the assumption made about the probability distribution of the residuals.

(7c) The sensitivity of the conclusions to the modeling assumptions, and model uncertainty in particular, should be evaluated by comparing results corresponding to different but similarly plausible models for the data (Clyde and George, 2004).

> EXAMPLE: Example E35 illustrates the evaluation of model uncertainty and uncertainty reduction by model averaging.

(7d) The statistical methods preferred in applications involving observation equations are likelihood-based, including maximum likelihood estimation and Bayesian procedures (DeGroot and Schervish, 2011; Wasserman, 2004), but *ad hoc* methods may be employed for special purposes (Example E22).

> EXAMPLES: Examples E2, E14, E17, E18, E20 and E27 illustrate maximum likelihood estimation and the corresponding evaluation of measurement uncertainty. Examples E19, E10, E22, E25, and E34 employ Bayesian procedures to estimate the measurand and to evaluate the associated measurement uncertainty.

(7e) When a Bayesian statistical procedure is employed to blend preexisting knowledge about the measurand or about the measurement procedure, with fresh experimental data, a so-called *prior* probability distribution must be assigned to the measurand that encapsulates that preexisting knowledge. In general, this distribution should be the result of a deliberate elicitation exercise that captures genuine prior knowledge (*cf.* Notes 5.2 and 5.3) rather than the result of applying formal rules (Kass and Wasserman, 1996).

(7f) In those rare cases where there is no credible *a priori* knowledge about the measurand but it is still desirable to employ a Bayesian procedure, then a so-called (non-informative) *reference prior* (Bernardo and Smith, 2007) may be used (Example E25).

8 Express measurement uncertainty in a manner that is fit for purpose. In most cases, specifying a set of values of the measurand believed to include its true value with 95 % probability (95 % coverage region) suffices as expression of measurement uncertainty.

(8a) When the result of an evaluation of measurement uncertainty is intended for use in subsequent uncertainty propagation exercises involving Monte Carlo 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 describes the state of knowledge about the measurand.

(8b) The techniques described in the GUM and in TN1297 produce approximate coverage intervals for scalar measurands. TN1297 (6.5) indicates that, by convention, the expanded uncertainty should be twice the standard uncertainty. This is motivated by the fact that, in many cases, a coverage interval of the form $y \pm 2u(y)$, where $u(y)$ denotes the standard uncertainty associated with y, achieves approximately 95 % coverage probability even when the probability distribution of the measurand is markedly skewed (that is, has one tail longer or heavier than the other) (Freedman, 2009).

However, TN1297 (Appendix B) also discusses when and how coverage intervals of the form $y \pm ku(y)$, with coverage factors k other than 2, may or should be used. Since the NUM implements the Monte Carlo method of the GUM-S1, it provides exact coverage intervals that will be symmetrical relative to y if symmetric intervals are requested, but that otherwise need not be symmetrical.

(8c) Coverage intervals or regions need not be symmetrical relative to the estimate of the measurand, and often the shortest or otherwise smallest such interval or region will not be symmetrical, especially when the measurand is constrained to be non-negative or to lie in a bounded region (Examples E3, E11, E19, E10, E25). In particular, unless explicitly instructed to produce symmetrical intervals, the NUM will often produce asymmetrical coverage intervals for scalar measurands.

> EXAMPLE: Asymmetric intervals are commonly used in nuclear physics. For example, Hosmer et al. (2010) reports the result of measuring the half-life of ^{80}Cu as 170^{+110}_{-50}ms.

(8d) An asymmetric coverage interval (for a scalar measurand) is defined by two numbers, $U_\gamma^-(y)$ and $U_\gamma^+(y)$ such that the interval from $y - U_\gamma^-(y)$ to $y + U_\gamma^+(y)$ is believed to include the true value of the measurand with a specified probability γ (which must be stated explicitly), typically 95 %.

(8e) When a symmetrical coverage interval with coverage probability $0 < \gamma < 1$ is desired for a scalar measurand (that is, an interval whose end-points are equidistant from the estimate of the measurand and that includes the true value of the measurand with probability γ), and the uncertainty evaluation was done using the Monte Carlo method, then determining such interval involves finding a positive number $U_\gamma(y)$ such that the interval $y \pm U_\gamma(y)$ includes a proportion γ of the values in the Monte Carlo sample, and leaves out the remaining $1 - \gamma$ proportion of the same sample. In such cases, the corresponding coverage factor is computed after the fact (*post hoc*) as $k = U_\gamma(y)/u(y)$, where $u(y)$ denotes the standard uncertainty associated with y, typically the standard deviation of the Monte Carlo sample that has been drawn from the probability distribution of y (Example E18).

NOTE 8.1 When it is desired to propagate the uncertainty expressed in an asymmetric coverage interval while preserving the asymmetry, a Monte Carlo method should be used, as illustrated in Example E3. For example, if the coverage probability is γ, then samples should be drawn from a probability distribution whose median (or, alternatively, whose mean) is equal to y and otherwise is such that it assigns probability γ to the interval from $y - U_\gamma^-(y)$ to $y + U_\gamma^+(y)$. In addition, this distribution should be generally consistent with the state of knowledge about the measurand.

NOTE 8.2 To propagate the uncertainty expressed in an asymmetric interval glossing over the asymmetry, define an approximate, effective standard uncertainty $u(y) = (U_{95\,\%}^-(y) + U_{95\,\%}^+(y))/4$, and use it in subsequent uncertainty propagation exercises. Audi et al. (2012, Appendix A) and Barlow (2003) describe other symmetrization techniques.

Examples

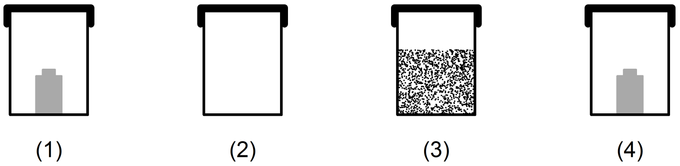
E1 Weighing. The mass m_P of a powder in a plastic container is measured using a single-pan electronic balance whose performance is comparable to the Mettler-Toledo XSE104 analytical balance, by taking the following steps:

(1) Determine the mass $c_{\mathrm{R},1}$ of a reference container that is nominally identical to the container with the powder, and contains a weight of mass 25 g, believed to be about half-way between the masses of the container with the powder and of the empty container;

(2) Determine the mass c_E of an empty container nominally identical to the container with the powder;

(3) Determine the mass c_P of the container with the powder;

(4) Determine the mass $c_{\mathrm{R},2}$ of the same reference container with the same weight inside that was weighed in the first step.



This procedure is a variant of weighing by differences, except that two nominally identical containers are being compared (one with the powder, the other empty), instead of weighing the same container before and after filling with the powder.

Notice that the weighing procedure involves three distinct, nominally identical containers. The container with the 25 g weight is weighed twice. Since the containers are weighed with tightly fitting lids on, and assuming that they all displace essentially the same volume of air and that the density of air remained essentially constant in the course of the weighings, there is no need for buoyancy corrections.

The masses of the nominally identical (empty) containers are known to have standard uncertainty 0.005 g. According to the manufacturer's specifications, the uncertainty of the weighings produced by the balance includes contributions from four sources of uncertainty related to the balance's performance attributes: readability ($u_\mathrm{B} = 0.1$ mg), repeatability ($u_\mathrm{R} = 0.1$ mg), deviation from linearity ($u_\mathrm{L} = 0.2$ mg), and eccentricity ($u_\mathrm{T} = 0.3$ mg), where the values between parentheses are the corresponding standard uncertainties.

The measurement equation is $m_\mathrm{P} = c_\mathrm{P} - c_\mathrm{E} - (c_{\mathrm{R},2} - c_{\mathrm{R},1})$: it expresses the output quantity m_P as a linear combination of the input quantities, which appear on the right-hand side. The second term on the right-hand side, $c_{\mathrm{R},2} - c_{\mathrm{R},1}$, the difference of the two weighings of the reference container, is intended to correct for temporal drift of the balance (Davidson et al., 2004, 3.2).

The uncertainties associated with the input quantities may be propagated to the output quantity in any one of at least three different ways: a method from the theory of probability, the method of the GUM (and of TN1297), or the Monte Carlo method of the GUM-S1.

Probability Theory. The variance (squared standard deviation) of a sum or difference of uncorrelated random variables is equal to the sum of the variances of these random variables. Assuming that the weighings are uncorrelated, we have $u^2(m_P) = u^2(c_P) + u^2(c_E) + u^2(c_{R,2} - c_{R,1})$ exactly. Below it will become clear why we evaluate the uncertainty associated with the difference $c_{R,2} - c_{R,1}$ instead of the uncertainties associated with $c_{R,2}$ and $c_{R,1}$ individually.

We model these quantities as random variables because there is some uncertainty about their true values, and this *Simple Guide* takes the position that all property values that are known incompletely or imperfectly are modeled as random variables whose probability distributions describe the metrologist's state of knowledge about their true values (*cf.* "Orientation" on Page 1).

Now we make the additional assumption that the "errors" that affect each weighing and that are attributable to lack of readability and repeatability, and to deviations from linearity and eccentricity of the balance, also are uncorrelated. In these circumstances, $u^2(c_P) = u^2(c_E) = 0.005^2\,\text{g}^2 + u_B^2 + u_R^2 + u_L^2 + u_T^2 = (0.005015)^2\,\text{g}^2$.

Concerning $u^2(c_{R,2} - c_{R,1})$: since $c_{R,2} - c_{R,1}$ is the difference between two weighings of the same container with the same weight inside, neither the uncertainty associated with the mass of the container, nor the uncertainty associated with the mass of the weight that it has inside, contribute to the uncertainty of the difference. Therefore, $u^2(c_{R,2} - c_{R,1}) = 2(u_B^2 + u_R^2 + u_L^2 + u_T^2) = (0.0005477)^2\,\text{g}^2$. Only the performance characteristics of the balance contribute to this uncertainty. The factor 2 is there because two weighings were made.

If the results of the four weighings are $c_P = 53.768\,\text{g}$, $c_E = 3.436\,\text{g}$, $c_{R,1} = 3.428\,\text{g}$, and $c_{R,2} = 3.476\,\text{g}$, we conclude that $m_P = 50.284\,\text{g}$ with associated standard uncertainty $u(m_P) = ((0.005015)^2 + (0.005015)^2 + (0.0005477)^2)^{½}\,\text{g} = 0.0071\,\text{g}$.

If we assume further that the input quantities are Gaussian random variables, a result from probability theory (the sum of independent Gaussian random variables is a Gaussian random variable), implies that the uncertainty associated with m_P is described fully by a Gaussian distribution with mean 50.284 g and standard deviation $u(m_P) = 0.007\,\text{g}$, hence the interval $m_P \pm 1.96u(m_P)$, which ranges from 50.270 g to 50.298 g, is a 95 % coverage interval for m_P. (The coverage factor 1.96 is the 97.5th percentile of a Gaussian distribution with mean 0 and standard deviation 1, to achieve 95 % coverage: in practice it is often rounded to 2.)

GUM & Monte Carlo. To use the NUM we regard the output quantity $y = m_P$ as a function of three input quantities: $x_1 = c_P$, $x_2 = c_E$ and $x_3 = c_{R,2} - c_{R,1}$, with $x_1 = 53.768\,\text{g}$, $u(x_1) = 0.005\,015\,\text{g}$, $x_2 = 3.436\,\text{g}$, $u(x_1) = 0.005\,015\,\text{g}$, $x_3 = 0.048\,\text{g}$, and $u(x_3) = 0.000\,547\,7\,\text{g}$. When Gaussian distributions are assigned to these three inputs, with means and standard deviations set equal to these estimates and standard uncertainties, both sets of results (GUM and Monte Carlo) produced by the NUM reproduce the results above.

In this case, because the output quantity is a linear combination of the input quantities, the approximation to $u(m_P)$ that the NUM produces when using Gauss's formula (Equation (A-3) of TN 1297 and Equation (13) in the GUM) is exact.

The Monte Carlo method may well be the most intuitive way of propagating uncertainties. It involves drawing one value x_1^* from the distribution of the first input, one value x_2^* from the distribution of the second input, one value x_3^* from the distribution of the third input, and then computing a value $y^* = x_1^* - x_2^* - x_3^*$ from the distribution of the output. Repeating this process many times produces a sample from the probability distribution of m_P, whose standard deviation is the evaluation of $u(m_P)$ according to the Monte Carlo method described by Morgan and Henrion (1992) and in the GUM-S1.

Since 95 % of the sample values lay between 50.270 g and 50.298 g, these are the endpoints of a 95 % coverage interval for the true value of m_P. Exhibit 1 shows a smooth histogram of these values, and also depicts the estimate of m_P and this coverage interval.

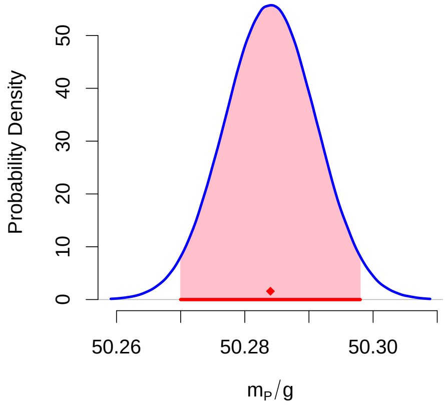


Exhibit 1: Smooth histogram of the sample drawn from the probability distribution of m_Pn produced by the Monte Carlo method. The estimate of m_P is indicated by a (red) diamond), and the 95 % coverage interval for the true value of m_P is represented by a thick, horizontal (red) line segment. The shaded (pink) region comprises 95 % of the area under the curve and above the horizontal line at ordinate 0.

E2 Surface Temperature. Exhibit 2 lists and depicts the values of the daily maximum temperature that were observed on twenty-two (non-consecutive) days of the month of May, 2012, using a traditional mercury-in-glass "maximum" thermometer located in the Stevenson shelter in the NIST campus that lies closest to interstate highway I-270.

DAY	1	2	3	4	7	8	9	10	11	14	15
$t/°C$	18.75	28.25	25.75	28.00	28.50	20.75	21.00	22.75	18.50	27.25	20.75
DAY	16	17	18	21	22	23	24	25	29	30	31
$t/°C$	26.50	28.00	23.25	28.00	21.75	26.00	26.50	28.00	33.25	32.00	29.50

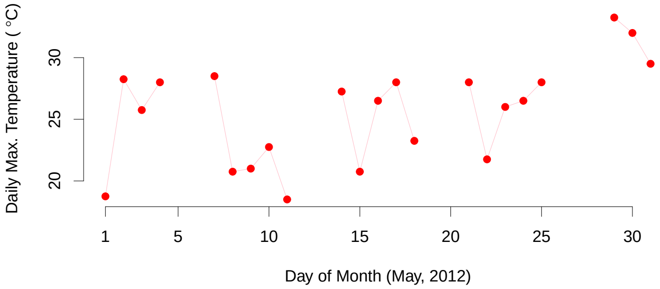


Exhibit 2: Values of daily maximum temperature measured during the month of May, 2012, using a mercury-in-glass "maximum" thermometer mounted inside a Stevenson shelter compliant with World Meteorological Organization guidelines (World Meteorological Organization, 2008, Chapter 2), deployed in the NIST Gaithersburg campus.

The average $\bar{t} = 25.59$ °C of these readings is a commonly used estimate of the daily maximum temperature τ during that month. The adequacy of this choice is contingent on the definition of τ and on a model that explains the relationship between the thermometer readings and τ.

The daily maximum temperature τ in the month of May, 2012, in this Stevenson shelter, may be defined as the mean of the thirty-one true daily maxima of that month in that shelter. The daily maximum t_i read on day $1 \leqslant i \leqslant 31$ typically deviates from τ owing to several effects, some of them persistent, affecting all the observations similarly, others volatile. Among the persistent effects there is possibly imperfect calibration of the thermometer. Examples of volatile effects include operator reading errors.

If ε_i denotes the combined result of such effects, then $t_i = \tau + \varepsilon_i$ where ε_i denotes a random variable with mean 0, for $i = 1, \dots, m$, where $m = 22$ denotes the number of days in which the thermometer was read. This so-called measurement error model (Freedman et al., 2007) may be specialized further by assuming that $\varepsilon_1, \dots, \varepsilon_m$ are modeled independent random variables with the same Gaussian distribution with mean 0 and standard deviation σ. In these circumstances, the $\{t_i\}$ will be like a sample from a Gaussian distribution with mean τ and standard deviation σ (both unknown).

The assumption of independence may obviously be questioned, but with such scant data it is difficult to evaluate its adequacy (Example E20 describes a situation where dependence is obvious and is taken into account). The assumption of Gaussian shape may be evaluated using a statistical test. For example, in this case the test suggested by Anderson and Darling (1952) offers no reason to doubt the adequacy of this assumption. However, because the dataset is quite small, the test may have little power to detect a violation of the assumption.

The equation, $t_i = \tau + \varepsilon_i$, that links the data to the measurand, together with the assumptions made about the quantities that figure in it, is the observation equation. The measurand τ is a parameter (the mean in this case) of the probability distribution being entertained for the observations.

Adoption of this model still does not imply that τ should be estimated by the average of the observations — some additional criterion is needed. In this case, several well-known and widely used criteria do lead to the average as "optimal" choice in one sense or another: these include maximum likelihood, some forms of Bayesian estimation, and minimum mean squared error.

The associated uncertainty depends on the sources of uncertainty that are recognized, and on how their individual contributions are evaluated.

One potential source of uncertainty is model selection: in fact, and as already mentioned, a model that allows for temporal correlations between the observations may very well afford a more faithful representation of the variability in the data than the model above. However, with as few observations as are available in this case, it would be difficult to justify adopting such a model.

The $\{\varepsilon_i\}$ capture three sources of uncertainty: natural variability of temperature from day to day, variability attributable to differences in the time of day when the thermometer was read, and the components of uncertainty associated with the calibration of the thermometer and with reading the scale inscribed on the thermometer.

Assuming that the calibration uncertainty is negligible by comparison with the other uncertainty components, and that no other significant sources of uncertainty are in play, then the common end-point of several alternative analyses is a scaled and shifted Student's t distribution as full characterization of the uncertainty associated with τ.

For example, proceeding as in the GUM (4.2.3, 4.4.3, G.3.2), the average of the $m = 22$ daily readings is $\bar{t} = 25.6$ °C, and the standard deviation is $s = 4.1$ °C. Therefore, the standard uncertainty associated with the average is $u(\tau) = s/\sqrt{m} = 0.872$ °C. The coverage factor for 95 % coverage probability is $k = 2.08$, which is the 97.5th percentile of Student's t distribution with 21 degrees of freedom. In this conformity, the shortest 95 % coverage interval is $\bar{t} \pm ks/\sqrt{n} = (23.8\,°\mathrm{C}, 27.4\,°\mathrm{C})$.

A coverage interval may also be built that does not depend on the assumption that the data are like a sample from a Gaussian distribution. The procedure developed by Frank Wilcoxon in 1945 produces an interval ranging from 23.6 °C to 27.6 °C (Wilcoxon, 1945; Hollander and Wolfe, 1999). The wider interval is the price one pays for no longer relying on any specific assumption about the distribution of the data.

E3 Falling Ball Viscometer. The dynamic viscosity μ_{M} of a solution of sodium hydroxide in water at 20 °C, is measured using a boron silica glass ball of mass density $\rho_{\mathrm{B}} = 2217\,\mathrm{kg/m^3}$. The measurement equation is $\mu_{\mathrm{M}} = \mu_{\mathrm{C}}\,[(\rho_{\mathrm{B}} - \rho_{\mathrm{M}})/(\rho_{\mathrm{B}} - \rho_{\mathrm{C}})]\,(t_{\mathrm{M}}/t_{\mathrm{C}})$, where $\mu_{\mathrm{C}} = 4.63\,\mathrm{mPa\,s}$, $\rho_{\mathrm{C}} = 810\,\mathrm{kg/m^3}$, and $t_{\mathrm{C}} = 36.6\,\mathrm{s}$ denote the viscosity, mass density, and ball travel time for the calibration liquid, and $\rho_{\mathrm{M}} = 1180\,\mathrm{kg/m^3}$ and $t_{\mathrm{M}} = 61\,\mathrm{s}$ denote the mass density and ball travel time for the sodium hydroxide solution (Exhibit 3).

If the input quantities are modeled as independent Gaussian random variables with means equal to their assigned values, and standard deviations equal to their associated standard uncertainties $u(\mu_{\mathrm{C}}) = 0.01\mu_{\mathrm{C}}$, $u(\rho_{\mathrm{B}}) = u(\rho_{\mathrm{C}}) = u(\rho_{\mathrm{M}}) = 0.5\,\mathrm{kg/m^3}$, $u(t_{\mathrm{C}}) = 0.15t_{\mathrm{C}}$, and $u(t_{\mathrm{M}}) = 0.10t_{\mathrm{M}}$, then the Monte Carlo method of the GUM-S1 as implemented in the NUM produces: $\mu_{\mathrm{M}} = 5.82\,\mathrm{mPa\,s}$ and $u(\mu_{\mathrm{M}}) = 1.11\,\mathrm{mPa\,s}$. The interval from 4.05 mPa s to 8.39 mPa s is an approximate 95 % coverage interval for μ_{M}, which happens to be asymmetric relative to the measured value.

Note that several of the standard uncertainties quoted may be unrealistically large for state-of-the-art laboratory practices, in particular for the ball travel times. These values have been selected to enhance several features of the results that otherwise might not stand out as clearly and that should be noted.

If the estimates of the input quantities had been substituted into the measurement equation as the GUM suggests, the resulting estimate of μ_{M} would have been 5.69 mPa s. And if the conventional formula for uncertainty propagation (Equation (A-3) of TN 1297 and Equation (13) in the GUM), which also is implemented in the NUM, had been used to evaluate $u(\mu_{\mathrm{M}})$, then the result would have been 1.11 mPa s.

Interestingly, the evaluation of $u(\mu_{\mathrm{M}})$ is identical to the evaluation produced by the Monte Carlo method, but the estimates of the measurand produced by one and by the other differ. Exhibit 3 shows that the coverage interval given above differs from the interval corresponding to the prescription in Clause 6.2.1 of the GUM (estimate of the output quantity plus or minus twice the standard measurement uncertainty evaluated using the approximate propagation of error formula). The difference is attributable to the skewness (or, asymmetry) of the distribution of the measurand, with a right tail that is longer (or, heavier) than the left tail.

If the Monte Carlo sample were no longer available, and the results of the uncertainty evaluation had been expressed only by specifying the asymmetrical 95 % coverage interval given above, ranging from 4.05 mPa s to 8.39 mPa s, and there was a need to propagate this uncertainty further, then the guidance offered in (8c) may be implemented as follows:

- Find a gamma probability distribution whose median equals the measured value, 5.82 mPa s, and otherwise is such that it assigns probability 95 % to the interval from 4.05 mPa s to 8.39 mPa s;
- Draw a sufficiently large sample from this distribution to be used in the subsequent Monte Carlo uncertainty propagation exercise.

Finding such gamma probability distribution can be accomplished by numerical minimization of the function that at α and β takes the value $(F_{\alpha,\beta}(8.39) - F_{\alpha,\beta}(4.05) - 0.95)^2 + (F_{\alpha,\beta}(5.82) - 0.5)^2$, where $F_{\alpha,\beta}$ denotes the cumulative probability distribution function of the gamma distribution with shape α and scale β. One solution of this minimization problem is $\widehat{\alpha} = 29.48$ and $\widehat{\beta} = 0.1997$ mPa s. The corresponding probability density is depicted in Exhibit 3.

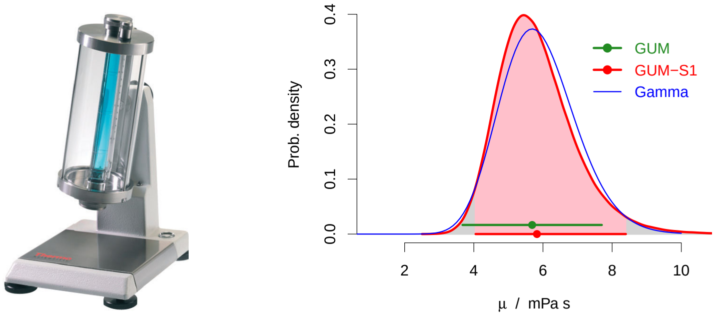


Exhibit 3: HAAKE™ falling ball viscometer from Thermo Fisher Scientific, Inc., (left panel), and probability density (right panel) corresponding to a Monte Carlo sample of size 1×10^6, also showing 95 % coverage intervals for the value of the dynamic viscosity of the liquid, one corresponding to the prescription in Clause 6.2.1 of the GUM, the other whose endpoints are the 2.5th and 97.5th percentiles of the Monte Carlo sample. The thin (blue) curve is the probability density of the gamma distribution with median equal to the estimate of the measurand, and 2.5th and 97.5th percentiles equal to the corresponding percentiles of the Monte Carlo sample.

E4 Pitot Tube. The pioneering work of Kline and McClintock (1953) predates the GUM by more than forty years but already includes all the key concepts elaborated in the GUM: (i) recognition that “in most engineering experiments it is not practical to estimate all of the uncertainties of observation by repetition”; (ii) measurement uncertainty should be characterized probabilistically; (iii) errors of different kinds (in particular “fixed” and “accidental” errors) should be described in the same manner, that is, via probability distributions that characterize uncertainty (“uncertainty distributions”), and should receive equal treatment; (iv) intervals qualified by odds (of including the true value of the measurand) are useful summaries of uncertainty distributions; and (v) uncertainty propagation, from inputs to output, may be carried out approximately using the formula introduced by Gauss (1823) that became Equation (10) in the GUM.

A typical Pitot tube used to measure airspeed has an orifice facing directly into the air flow to measure total pressure, and at least one orifice whose surface normal is orthogonal to the flow to measure static pressure (Exhibit 4). Airspeed v is determined by the difference Δ between the total and static pressures, and by the mass density ρ of air, according to the measurement equation $v = \sqrt{2\Delta/\rho}$. Since ρ is usually estimated by application of the ideal gas law, the measurement equation becomes $v = \sqrt{2\Delta R_s T/p}$, where p and T denote the air pressure and temperature, and $R_s = 287.058\,\mathrm{J\,kg^{-1}\,K^{-1}}$ is the specific gas constant for dry air.

Exhibit 4: Pitot tube mounted on a helicopter (Zátonyi Sándor, `en.wikipedia.org/wiki/Pitot_tube`) showing one large, forward-facing, circular orifice to measure total pressure, and several small circular orifices behind a trim ring, to measure static pressure.

Kline and McClintock (1953) illustrate the method to evaluate the uncertainty associated with v in a case where $\Delta = 1.993\,\mathrm{kPa}$ was measured with a U-tube manometer, $p = 101.4\,\mathrm{kPa}$ was measured with a Bourdon gage, and $T = 292.8\,\mathrm{K}$ was measured with a mercury-in-glass thermometer. The expanded uncertainties (which they characterize as 95 % coverage intervals by saying that they are defined "with odds of 20 to 1") were $U_{95\,\%}(\Delta) = 0.025\,\mathrm{kPa}$, $U_{95\,\%}(p) = 2.1\,\mathrm{kPa}$, and $U_{95\,\%}(T) = 0.11\,\mathrm{K}$. (The original treatment disregards the uncertainty component affecting R_s that is attributable to lack of knowledge about the actual humidity of air.)

Taking the corresponding standard uncertainties as one half of these expanded uncertainties, the NUM produces $v = 40.64\,\mathrm{m/s}$ and $u(v) = 0.25\,\mathrm{m/s}$ according to both Gauss's formula and the Monte Carlo method (for which the input variables were modeled as Gaussian random variables). An approximate 95 % coverage interval defined as $v \pm 2u(v)$ ranges from 40.15 m/s to 41.14 m/s. Its counterpart based on the results of the Monte Carlo method, with endpoints given by the 2.5th and 97.5th percentiles of a sample of size 1×10^6 drawn from the distribution of v, ranges from 40.17 m/s to 41.13 m/s.

E5 Gauge Blocks. Exhibit 5 shows a single-probe mechanical comparator used to measure dimensions of gauge blocks by comparison with dimensions of master blocks, as described by Doiron and Beers (1995, Section 5.4).

The measurement involves: (i) obtaining the readings x and r that the comparator produces when presented with the block that is the target of measurement and with a reference block of the same nominal length, (ii) applying a correction for the difference in deformation between the two blocks that is caused by the force that the probe makes while in contact with their surfaces, (iii) applying a correction that accounts for the difference between the thermal expansion coefficients of the blocks and also for the difference between the ambient temperature and the reference temperature of 20 °C, and (iv) characterizing and propagating the uncertainties associated with the inputs.

The measurement equation is $L_x = L_r + (x - r) + (\delta_x - \delta_r) + L(\alpha_r - \alpha_x)(t - 20)$ (Doiron and Beers, 1995, Equation (5.4)), where L_x and L_r denote the lengths of the measured and

Exhibit 5: Version of the Mahr-Federal comparator model 130B-24 used by the Dimensional Metrology Group (Semiconductor and Dimensional Metrology Division, Physical Measurement Laboratory, NIST) for the mechanical comparison of dimensions of gauge blocks.

reference blocks, δ_x and δ_r denote the elastic deformations induced by the force that the probe exerts upon the surfaces of the blocks, L denotes the common nominal length of the blocks, α_r and α_x denote their thermal expansion coefficients, and t denotes the temperature of the environment that the blocks are assumed to be in thermal equilibrium with during measurement.

A tungsten carbide block of nominal length $L = 50\,\text{mm}$ was measured using a steel block of the same nominal length as reference, whose actual length was $L_r = 50.000\,60\,\text{mm}$. The comparator readings were $x = 1.25 \times 10^{-3}\,\text{mm}$ for the tungsten carbide block, and $r = 1.06 \times 10^{-3}\,\text{mm}$ for the reference steel block (Doiron and Beers, 1995, 5.3.1, Example 1).

The corresponding thermal expansion coefficients were $\alpha_x = 6 \times 10^{-6}\,°\text{C}^{-1}$ and $\alpha_r = 11.5 \times 10^{-6}\,°\text{C}^{-1}$. The contact deformations, corresponding to a force of 0.75 N applied by the probe onto the surface of the blocks, are estimated as $\delta_x = 0.08 \times 10^{-3}\,\text{mm}$ for the block being measured, and $\delta_r = 0.14 \times 10^{-3}\,\text{mm}$ and for the reference block (Doiron and Beers, 1995, Table 3.4). The ambient temperature was $t = 20.4\,°\text{C}$.

Therefore, $L_x = 50.00060 + (1.25 - 1.06) \times 10^{-3} + (0.08 - 0.14) \times 10^{-3} + 50 \times (11.5 - 6) \times 10^{-6} \times (20.4 - 20) = 50.000\,84\,\text{mm}$. The associated uncertainty is evaluated by propagating the contributions recognized in Exhibit 6.

Doiron and Stoup (1997) point out that the uncertainty associated with the coefficient of thermal expansion depends on the length of the block because in steel blocks at least, the value of the coefficient varies between the ends of the blocks (where the steel has been hardened), and their central portions (which remain unhardened).

For the NUM to be able to recognize the contributions that scale calibration (S), instrument geometry (I), and artifact geometry (A) make to the overall measurement uncertainty, input quantities need be introduced explicitly whose estimated values are zero but whose standard uncertainties are as listed in Exhibit 6. In consequence, the measurement equation becomes $L_x = L_r + (x - r) + (\delta_x - \delta_r) + L(\alpha_r - \alpha_x)(t - 20) + S + I + A$ where S, I, and A are estimated as 0, with $u(S) = 0.002 \times 10^{-3}\,\text{mm}$, $u(I) = 0.002 \times 10^{-3}\,\text{mm}$, and $u(A) = 0.008 \times 10^{-3}\,\text{mm}$. The nominal length L of the blocks is treated as a known constant.

Application of Gauss's formula as implemented in the NUM produces the estimate $L_x = 50.000\,84\,\text{mm}$ and $u(L_x) = 1.6 \times 10^{-5}\,\text{mm}$. For the Monte Carlo method, L_r, x, r, S, I, and A are modeled as Gaussian random variables with means and standard deviations set equal to their estimates and standard uncertainties; α_x, α_r, and t are modeled as random variables with uniform (or, rectangular) distributions; and δ_x and δ_r are modeled as random variables with Gaussian distributions truncated at zero. These random variables are assumed

SOURCE	STANDARD UNCERTAINTY ($k = 1$)
Master Gauge Calibration	0.012×10^{-3} mm + ($L \times 0.08 \times 10^{-9}$)
Reproducibility	0.004×10^{-3} mm + ($L \times 0.12 \times 10^{-9}$)
Coeff. of Thermal Expansion	$L \times 0.20 \times 10^{-9}$
Thermal Gradients	$L \times 0.17 \times 10^{-9}$
Contact Deformation	0.002×10^{-3} mm
Scale Calibration	0.002×10^{-3} mm
Instrument Geometry	0.002×10^{-3} mm
Artifact Geometry	0.008×10^{-3} mm

Exhibit 6: Uncertainty budget as specified by Doiron and Stoup (1997, Table 6), except for the coefficient of thermal expansion, whose standard uncertainty is as listed in Doiron and Beers (1995, Table 4.3), where L denotes the nominal length of the block, expressed in millimeter.

to be mutually independent.

A sample of size 10^6 drawn from the probability distribution of L_x had mean 50.000 84 mm and standard deviation $u(L_x) = 1.6 \times 10^{-5}$ mm. A 95 % symmetrical coverage interval for the true value of L_x, computed directly from the Monte Carlo sample, ranges from 50.000 81 mm to 50.000 87 mm. The corresponding expanded uncertainty is $U_{95\,\%}(L_x) = 3.1 \times 10^{-5}$ mm.

E6 DNA Sequencing. The first measurand θ to be considered is a finite sequence of letters that represent the identities of the nucleobases (A for adenine, C for cytosine, G for guanine, and T thymine) along a fragment of a strand of deoxyribonucleic acid (DNA). The sequencing procedure yields an estimate of this measurand, say

$$\widehat{\theta} = (\text{TTTTTATAATTGGTTAATCATTTTTTTTTAATTTTT}).$$

Some sequencing techniques compute the probability of the nucleobase at any given location being A, C, G, or T, and then assign to the location the nucleobase that has the highest probability. These probabilities are often represented by integer *quality scores*. For example, the line for location 7 in Exhibit 8 lists the scores assigned to the four bases: $Q(\text{A}) = -14$, $Q(\text{C}) = -10$, $Q(\text{G}) = -12$, and $Q(\text{T}) = 6$. The larger the score, the greater the confidence in the corresponding base as being the correct assignment to that location: T in this case.

These scores are of the form $Q = -10\log_{10}(e/(1-e))$, where e denotes the probability of error if the corresponding base is assigned to the location. For location 7, $Q(\text{A}) = -14$, which means that the odds against A at this location are $o = e/(1-e) = 10^{1.4} = 25$, or, equivalently, that the probability of A at this location is $\Pr(\text{A}) = 1/(1+o) = 0.04$.

Therefore, the quadruplet ($\Pr(\text{A})$, $\Pr(\text{C})$, $\Pr(\text{G})$, $\Pr(\text{T})$) associated with each location is a probability distribution over the set of possible values $\{\text{A}, \text{C}, \text{G}, \text{T}\}$. These probability distributions (one for each location) characterize measurement uncertainty fully, and also suggest which nucleobase should be assigned to each location. For example, for location 7, $\Pr(\text{A}) = 0.04$, $\Pr(\text{C}) = 0.09$, $\Pr(\text{G}) = 0.06$, and $\Pr(\text{T}) = 0.81$, and T was identity assigned to this location because it has the largest probability.

The implied dispersion of values (of the nominal property that is the identity of the base) may be summarized by the entropy of this distribution, $H = -\Pr(\mathrm{A})\log\Pr(\mathrm{A}) - \Pr(\mathrm{C})\log\Pr(\mathrm{C}) - \Pr(\mathrm{G})\log\Pr(\mathrm{G}) - \Pr(\mathrm{T})\log\Pr(\mathrm{T})$. For example, for location 5 the entropy is 0.07, while for location 7 it is 0.69, consistently with the perception that the distribution is much more concentrated for the former than for the latter. The values of H are listed in Exhibit 8, but they are not otherwise used in this example.

The uncertainty associated with each base call may be propagated to derivative quantities. Consider, as our second measurand, the Damerau-Levenshtein distance $D(\theta, \tau)$ (Damerau, 1964; Levenshtein, 1966) between the measurand θ and the following target sequence, which could be a known gene that θ is being compared against:

$$\tau = (\mathrm{GGATTTTATTATAAATGGGTATACAATTTTTAAAATTTT}).$$

Since $D(\theta, \tau)$ is the minimum number of insertions, deletions, or substitutions of a single character, or transpositions of two adjacent characters that are needed to transform one string into the other, θ and τ may very well have different lengths, as they do in this case. $D(\theta, \tau)$ is estimated as $D(\widehat{\theta}, \tau) = 13$, where D is evaluated using function `stringdist` defined in the R package of the same name (van der Loo, 2014). Exhibit 7 describes the 13 steps that lead from $\widehat{\theta}$ to τ.

$\widehat{\theta}$	TTT TTATAATTGGTTAATCATTTTTTTTTAATTTTT	
1	**G** TTT TTATAATTGGTTAATCATTTTTTTTTAATTTTT	
2	G**G** TTT TTATAATTGGTTAATCATTTTTTTTTAATTTTT	
3	GG**A** TTT TTATAATTGGTTAATCATTTTTTTTTAATTTTT	
4	GGA**T**TTT TTATAATTGGTTAATCATTTTTTTTTAATTTTT	
5	GGATTTT**A**TTATAATTGGTTAATCATTTTTTTTTAATTTTT	
6	GGATTTTATTATAA**A**TGGTTAATCATTTTTTTTTAATTTTT	
7	GGATTTTATTATAAATGG**G**TAATCATTTTTTTTTAATTTTT	
8	GGATTTTATTATAAATGGGTA**TA**CATTTTTTTTTAATTTTT	
9	GGATTTTATTATAAATGGGTATACA TTTTTTTTAATTTTT	
10	GGATTTTATTATAAATGGGTATACA TTTTTTTAATTTTT	
11	GGATTTTATTATAAATGGGTATACA **A**TTTTTTAATTTTT	
12	GGATTTTATTATAAATGGGTATACA ATTTTT**A**AATTTTT	
13	GGATTTTATTATAAATGGGTATACA ATTTTTAAA**A**TTTT	τ

Exhibit 7: Sequence of 13 steps (insertions, deletions, substitutions, or transpositions of two adjacent characters) that transform $\widehat{\theta}$ into τ.

To characterize the associated uncertainty, employ the Monte Carlo method:

1. Select a suitably large sample size K;
2. For each $k = 1, \dots, K$, and for each row of Exhibit 8, draw a letter from $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ using the probabilities in the same line, finally to obtain a string θ^*_k whose characters represent the nucleobases assigned to the thirty-six locations;
3. The distances $D(\theta^*_1, \tau), \dots, D(\theta^*_K, \tau)$ are a sample from the distribution of $D(\theta, \tau)$.

Exhibit 9 shows an estimate of the probability density of $D(\theta, \tau)$ based on a sample of size $K = 1 \times 10^5$, and it shows that $D(\theta, \tau) = 15$ is the value with highest probability, not the value (13) that was measured. In fact, the sample has average 15.2 and standard deviation $u(D(\theta, \tau)) = 1.4$, and a 95 % coverage interval for $D(\theta, \tau)$ ranges from 13 to 18.

LOC	Q(A)	Q(C)	Q(G)	Q(T)	BASE	Pr(A)	Pr(C)	Pr(G)	Pr(T)	H
1	−40	−40	−40	40	T	0.00	0.00	0.00	1.00	0.00
2	−40	−40	−40	40	T	0.00	0.00	0.00	1.00	0.00
3	−40	−40	−2	2	T	0.00	0.00	0.39	0.61	0.67
4	−21	−3	−40	3	T	0.01	0.33	0.00	0.66	0.68
5	−40	−19	−40	19	T	0.00	0.01	0.00	0.99	0.07
6	5	−40	−18	−5	A	0.75	0.00	0.02	0.24	0.62
7	−14	−10	−12	6	T	0.04	0.09	0.06	0.81	0.69
8	40	−40	−40	−40	A	1.00	0.00	0.00	0.00	0.00
9	9	−9	−40	−36	A	0.89	0.11	0.00	0.00	0.35
10	−13	−8	−40	6	T	0.05	0.14	0.00	0.81	0.60
11	−40	−40	−40	40	T	0.00	0.00	0.00	1.00	0.00
12	−1	−40	1	−32	G	0.44	0.00	0.56	0.00	0.69
13	−10	−30	8	−13	G	0.09	0.00	0.86	0.05	0.50
14	−10	−6	−4	−1	T	0.09	0.20	0.28	0.43	1.26
15	−25	−30	−40	24	T	0.00	0.00	0.00	1.00	0.03
16	40	−40	−40	−40	A	1.00	0.00	0.00	0.00	0.00
17	12	−14	−18	−34	A	0.95	0.04	0.02	0.00	0.25
18	−25	−40	−12	11	T	0.00	0.00	0.06	0.94	0.26
19	−15	12	−36	−15	C	0.03	0.94	0.00	0.03	0.27
20	12	−16	−40	−14	A	0.94	0.02	0.00	0.04	0.27
21	−28	−22	−40	21	T	0.00	0.01	0.00	0.99	0.05
22	−15	−29	−11	9	T	0.03	0.00	0.07	0.89	0.41
23	−40	−2	−40	2	T	0.00	0.39	0.00	0.61	0.67
24	−5	−10	−40	3	T	0.24	0.09	0.00	0.67	0.83
25	−24	−20	−10	10	T	0.00	0.01	0.09	0.90	0.37
26	−40	−31	−40	31	T	0.00	0.00	0.00	1.00	0.01
27	−40	−40	−40	40	T	0.00	0.00	0.00	1.00	0.00
28	−39	−40	−24	23	T	0.00	0.00	0.00	1.00	0.03
29	−40	−40	−40	40	T	0.00	0.00	0.00	1.00	0.00
30	1	−6	−14	−6	A	0.56	0.20	0.04	0.20	1.10
31	2	−11	−26	−4	A	0.63	0.08	0.00	0.29	0.86
32	−8	−40	−40	8	T	0.14	0.00	0.00	0.86	0.40
33	−1	−19	−11	−1	T	0.46	0.01	0.08	0.46	0.97
34	−29	−38	−40	29	T	0.00	0.00	0.00	1.00	0.01
35	−40	−40	−40	40	T	0.00	0.00	0.00	1.00	0.00
36	−40	−40	−40	40	T	0.00	0.00	0.00	1.00	0.00

Exhibit 8: DNA sequencing results from example `prb` and `seq` data files distributed with the `ShortRead` Bioconductor package for R (Morgan et al., 2009). Each line pertains to a location (LOC) in the sequence. Q(A), Q(C), Q(G), and Q(T) are the quality scores, and Pr(A), Pr(C), Pr(G), and Pr(T) are the corresponding probabilities. The values of the entropy of these discrete distributions are listed under H.

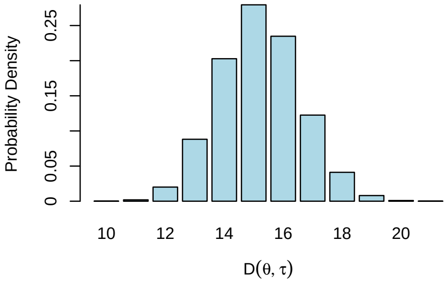


Exhibit 9: Estimate of the probability density of the Damerau-Levenshtein distance $D(\theta, \tau)$ between the measurand θ and the target sequence τ, derived from a sample of 1×10^5 replicates, whose mean and standard deviation were 15.2 and 1.4, while the measured value was 13.

E7 Thermistor Calibration. Whetstone et al. (1989) employed thermistor probes to measure the temperature of flowing water and of the atmosphere surrounding a weighing apparatus used to measure coefficients of discharge of orifice plates. These thermistors were calibrated by comparison with a platinum resistance thermometer (PRT) that had previously been calibrated by the Pressure and Temperature Division of what was then the National Bureau of Standards.

Calibration Data. Exhibit 10 lists the data used for the calibration of thermistor 775008 (Whetstone et al., 1989, Table 17), which comprise readings taken simultaneously with the thermistor and the PRT immersed in a thermostatically controlled bath filled with mineral oil.

	Temperature / °C											
PRT	20.91	25.42	30.50	34.96	40.23	34.93	30.05	25.03	20.87	16.41	16.40	39.34
THERMISTOR	20.85	25.52	30.70	35.22	40.47	35.18	30.25	25.10	20.81	16.23	16.22	39.56

Exhibit 10: Values of temperature of a thermostatically controlled bath measured simultaneously by a calibrated PRT and by thermistor 775008 (Whetstone et al., 1989, Table 17).

In many cases, there is a legal requirement for the calibration (Note 3.7) to characterize how the device being calibrated responds to the inputs that it is designed for. Here this means characterizing how the thermistor responds when immersed in a medium at the temperature indicated by the PRT that acts as a reference: we call the function that maps values of temperature indicated by the PRT to values of temperature indicated by the thermistor, the *calibration function*.

In practice, the thermistor will be used to measure the temperature of the medium it is immersed in and in thermal equilibrium with. This is done by reading the temperature that the thermistor produces, and then applying a function to it that produces a calibrated, measured temperature. This function is called the *analysis function*.

Calibration and Analysis Functions. The calibration and analysis functions often are mathematical inverses of one another. In this case, we will first build a function φ (calibration function) that expresses the indication I produced by the thermistor as a function of the temperature T measured by the PRT, $I = \varphi(T)$. But then, for practical use, we will require the inverse of this function, $\psi = \varphi^{-1}$ (analysis function), which maps the thermistor's reading into a traceable value of temperature, $T = \psi(I)$.

The nomenclature *calibration function* (φ) and *analysis function* ($\psi = \varphi^{-1}$) is used in ISO (2001) (Examples E17 and E18). The former characterizes how the thermistor responds to conditions that are essentially known (the temperature of the bath as measured by the PRT), while the latter predicts the true value of the temperature given a reading produced by the thermistor. The name *measurement function* may be a better name for the analysis function in a general context, but unfortunately it conflicts with how it is often used for the functions that appear in measurement equations.

The question may naturally be asked of why not build ψ directly, given that it is the function needed to use the thermistor in practice, instead of determining φ first, and then inverting it to obtain ψ.

One of the reasons has been indicated above already: the legal requirement for characterizing how the device being calibrated responds to known inputs. Another reason is that this more circuitous route, sometimes called *inverse regression* (Osborne, 1991), can be followed using conventional regression methods, instead of requiring more specialized software. (Examples E17 and E18 illustrate how the analysis function may be built directly.)

Both the indications provided by the thermistor and by the PRT are affected by errors. Determining a relationship between them involves finding a curve that minimizes the apparent errors and expresses one indication as a function of the other. In this case, and as will be seen shortly, the errors affecting the indications I provided by the thermistor are about 8 times larger than the errors affecting the temperatures T measured by the PRT. If the curve in question is to minimize deviations of the points (T, I) from it, and these deviations are measured either along the axis of temperatures T or along the axis of indications I, then it stands to reason that the curve should minimize the larger deviations, which in this case are between observed and predicted values of I.

In these circumstances, the statistical model underlying ordinary least squares regression is the observation equation $I_j = \varphi(T_j) + \varepsilon_j$, where $j = 1, \dots, n = 12$ identifies the set point at which temperature T_j was measured by the PRT during calibration, and the corresponding indication I_j was read off the thermistor, ε_j denotes the error affecting I_j, and T_j is assumed known without error, or at least known up to an error that is negligible by comparison with ε_j.

Model Selection and Fit. The calibration function φ will be a polynomial of low degree, and it will be fitted to the data by ordinary least squares, which is optimal (in several senses of "optimality") if the $\{\varepsilon_j\}$ are like a sample from a Gaussian distribution with mean 0.

The degree of the polynomial was selected by comparing polynomials of degrees from 1 to 6, using analysis of variance techniques (Chambers, 1991) that suggested a polynomial of the 3rd degree as representing the best compromise between goodness-of-fit and model parsimony: $I_j = \beta_0 + \beta_1 T_j + \beta_2 T_j^2 + \beta_3 T_j^3 + \varepsilon_j$ for $j = 1, \dots, n$.

The least squares estimates of the coefficients are $\widehat{\beta}_0 = -0.2785\ °\mathrm{C}$, $\widehat{\beta}_1 = 0.9722\ °\mathrm{C}^{-1}$, $\widehat{\beta}_2 = 0.002\,773\ °\mathrm{C}^{-2}$, and $\widehat{\beta}_3 = -4.404 \times 10^{-5}\ °\mathrm{C}^{-3}$. (These differ from the corresponding values in Whetstone et al. (1989, Table 18) because the latter pertain to a polynomial of the 3rd degree fitted to the $\{T_j\}$ as a function of the $\{I_j\}$). All except the intercept $\widehat{\beta}_0$ are statistically significantly different from 0.

Conventional graphical diagnostics — plot of residuals $\{\widehat{\varepsilon}_j\}$ against fitted values $\{\widehat{I}_j\}$, and QQ-plot of the residuals — reveal no obvious inadequacy of the model to these data. This calibration is valid for thermistor indications in the range $16.22\ °\mathrm{C} \leqslant I \leqslant 40.47\ °\mathrm{C}$.

Analysis Function. To find the calibrated value of temperature that corresponds to a reading I made by the thermistor involves solving the following equation for T: $\widehat{\beta}_0 + \widehat{\beta}_1 T + \widehat{\beta}_2 T^2 + \widehat{\beta}_3 T^3 = I$. Of the three solutions (generally complex numbers) that this equation will have, we select the one whose imaginary part is essentially equal to 0, and whose real part is between 16 °C and 40 °C, which is the calibration range.

For example, if $I = 27.68\ °\mathrm{C}$, computing $\psi(I)$ involves solving the cubic equation $-0.2785 + 0.9722T + 0.002773T^2 - 0.00004404T^3 = 27.68$. Of the three roots of this equation, 27.54 °C, −135.14 °C, and 170.57 °C, only the first is within the calibration range.

Exhibit 11 depicts the calibration and analysis functions, and also the expanded uncertainties associated with estimates of the temperature that the analysis function produces, evaluated as explained below.

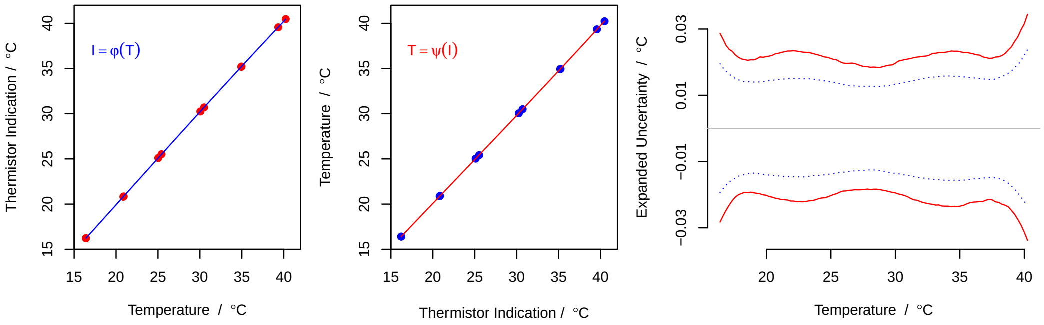


Exhibit 11: LEFT PANEL: Calibration function that produces the indication $I = \varphi(T)$ that the thermistor is expected to produce when the PRT indicates that the temperature is T. CENTER PANEL: Analysis function that produces the value $T = \psi(I)$ of temperature that corresponds to an indication I produced by the thermistor. The calibration and analysis functions, φ and $\psi = \varphi^{-1}$, appear to be identical only because I and T have the same units of measurement and corresponding values are numerically close to one another, and both functions are approximately linear over the ranges depicted. RIGHT PANEL: Simultaneous coverage envelopes for T, with coverage probabilities 68 % (dotted blue lines) and 95 % (solid red lines).

Uncertainty Evaluation. Whetstone et al. (1989, Page 62) reports that the standard uncertainty associated with the values of temperature measured by the PRT is $u_{\text{PRT}}(T) = 0.0015$ °C, and points out that extension cables used to connect the thermistor probe to the location where the indications were read also are a source of uncertainty with standard uncertainty $u_{\text{CABLE}}(T) = 0.01$ °C. These, and the contributions from the residuals $\{\widehat{\varepsilon}_j\}$, will be propagated using the Monte Carlo method, by taking the following steps:

1. Compute $\tau = \left(\widehat{\sigma}^2 + u^2_{\text{CABLE}}(T)\right)^{½} = 0.016$ °C, where $\widehat{\sigma} = 0.012$ °C is the estimate of the standard deviation of the residuals $\{\widehat{\varepsilon}_j\}$ corresponding to the polynomial fit for the calibration function.

2. Let $\theta_1, \dots, \theta_m$ denote a set of $m = 100$ values of indication values for the thermistor equispaced from 16.22 °C to 40.47 °C (these are the values at which the inverse of the calibration function will be evaluated for purposes of display as in Exhibit 11).

3. Choose a suitably large integer K (in this example $K = 10\,000$), and then for $k = 1, \dots, K$:

   (a) Draw $T_{1,k}, \dots, T_{n,k}$ independently from $n = 12$ Gaussian distributions with means $T_1, \dots, T_n$ (the values of temperature measured by the PRT) and standard deviations all equal to $u_{\text{PRT}}(T)$.

   (b) Draw $I_{1,k}, \dots, I_{n,k}$ independently from $n = 12$ Gaussian distributions with means $\widehat{I}_1, \dots, \widehat{I}_n$ (the thermistor indications predicted by the calibration function φ at the values of temperature measured by the PRT) and standard deviations all equal to τ.

(c) Determine the polynomial of the third degree φ^*_k by least squares that expresses the $\{I_{j,k} : j = 1, \dots, n\}$ as a function of the $\{T_{j,k} : j = 1, \dots, n\}$.

(d) For each $i = 1, \dots, m$, compute $T^*_{i,k} = \widehat{\psi}^*_k(\theta_i)$, where ψ^*_k denotes the inverse of φ^*_k. This step involves solving a cubic equation for each i, and determining the suitable root to assign to $T^*_{i,k}$.

4. Determine coverage intervals, depicted in Exhibit 11, for all values of $i = 1, \dots, m$ simultaneously, applying the method described by Davison and Hinkley (1997, Section 4.2.4) and implemented in R function `envelope` (Canty and Ripley, 2013b), using the data in the $m \times K$ array with the $\{T^*_{i,k}\}$.

E8 Molecular weight of carbon dioxide. The relative molecular mass (or, molecular weight) of carbon dioxide is $M_r(CO_2) = A_r(C) + 2A_r(O)$, where $A_r(C)$ and $A_r(O)$ denote the relative atomic masses (or, atomic weights) of carbon and oxygen.

The standard atomic weights of carbon and oxygen are intervals that describe the diversity of isotopic compositions of these elements in normal materials: $A_r(C) = [12.0096, 12.0116]$ and $A_r(O) = [15.999\,03, 15.999\,77]$ (Wieser et al., 2013).

The Commission on Isotopic Abundances and Atomic Weights (CIAAW) of the International Union of Pure and Applied Chemistry (IUPAC), defines *normal material* for a particular element, as any terrestrial material that “is a reasonably possible source for this element or its compounds in commerce, for industry or science; the material is not itself studied for some extraordinary anomaly and its isotopic composition has not been modified significantly in a geologically brief period” (Peiser et al., 1984).

If $A^*_r(C)$ and $A^*_r(O)$ denote independent random variables with uniform (or, rectangular) distributions over those intervals, then their mean values are 12.0106 and 15.9994 (which are the midpoints of the intervals), and their standard deviations are $u(A_r(C)) = 0.0006$ and $u(A_r(O)) = 0.0002$ (the standard deviation of a uniform distribution equals the length of the interval where the distribution is concentrated, divided by $\sqrt{12}$).

Therefore, $12.0106 + 2(15.9994) = 44.0094$ is an estimate of $M_r(CO_2)$. Neglecting the diminutive correlation between $A^*_r(C)$ and $A^*_r(O)$ that is induced by the implied normalization relative to the atomic mass of ^{12}C, $M_r(CO_2)$ is a linear combination of two uncorrelated random variables.

According to a result of probability theory, the variance of $M_r(CO_2)$ is equal to the variance of $A_r(C)$ plus 4 times the variance of $A_r(O)$: $u^2(M_r(CO_2)) = u^2(A_r(C)) + 4u^2(A_r(O)) = (0.0006)^2 + 4(0.0002)^2 = (0.000721)^2$. Therefore, $u(M_r(CO_2)) = 0.0007$. If either the Monte Carlo method used in Example E1, or the conventional error propagation formula of the GUM were used, the same results would have been obtained.

In this case it is also possible to derive analytically not only the standard uncertainty $u(M_r(CO_2))$, but the whole probability distribution that characterizes the uncertainty associated with the molecular weight of CO_2.

In fact, $M^*_r(CO_2) = A^*_r(C) + 2A^*_r(O)$ is a random variable with a symmetrical trapezoidal distribution with the mean and standard deviation given above, and whose probability density is depicted in Exhibit 12 (Killmann and von Collani, 2001). Using this fact, exact coverage intervals can be computed: for example, $[44.0080, 44.0108]$ is the shortest 95 % coverage interval for the molecular weight of carbon dioxide.

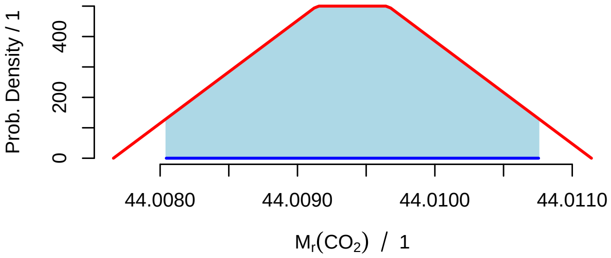


Exhibit 12: Trapezoidal probability density that characterizes the uncertainty associated with the molecular weight of carbon dioxide, assuming that the atomic weights of carbon and oxygen are independent random variables distributed uniformly over the corresponding standard atomic weight intervals. The shaded region comprises 95 % of the area under the trapezoid, and its footprint on the horizontal axis is the shortest, exact 95 % coverage interval.

E9 Cadmium Calibration Standard. A calibration standard for atomic absorption spectroscopy is prepared by adding a mass m of cadmium, with purity P, to an acidic solvent to obtain a solution of volume V (Ellison and Williams, 2012, Example A1). The measurement equation expresses the concentration of cadmium as $c_{\text{Cd}} = 1000mP/V$. The input quantities have the following values and standard uncertainties: $m = 100.28$ mg, $u(m) = 0.05$ mg; $P = 0.9999$, $u(P) = 0.000\,058$; $V = 100.0$ mL, $u(V) = 0.07$ mL. Therefore, $c_{\text{Cd}} = 1002.7$ mg/L.

If the goal is simply to compute an approximation to $u(c_{\text{Cd}})$, and given that the input quantities are combined using multiplications and divisions only, then, according to the GUM 5.1.6, the squared relative uncertainty of c_{Cd} is approximately equal to the sum of the squared relative uncertainties of the input quantities, $(u(c_{\text{Cd}})/c_{\text{Cd}})^2 \approx (u(m)/m)^2 + (u(P)/P)^2 + (u(V)/V)^2$, hence $u(c_{\text{Cd}}) \approx 0.9$ mg/L. The NUM reproduces this result.

In the absence of specific additional information about these quantities, to apply the Monte Carlo method we may assume that the corresponding random variables have Gaussian distributions with the means and standard deviations equal to the values and standard uncertainties given above. In these circumstances, the Monte Carlo method as implemented in the NUM produces results practically identical to those listed above.

The results are still the same if the models described in Ellison and Williams (2012, Example A1) for P (uniform distribution between 0.9998 and 1) and for V (symmetrical triangular distribution with mean 100 mL and standard deviation 0.7 mL) are used instead. A 95 % coverage interval based on a Monte Carlo sample of size 10^6 ranges from 1001.0 mg/L to 1004.4 mg/L.

E10 PCB in Sediment. Key Comparison CCQM–K25 was carried out to compare the results of the determination of the mass fractions of five different polychlorinated biphenyl (PCB) congeners in sediment (Schantz and Wise, 2004). Exhibit 13 lists and depicts the selected results for PCB 28 (2, 4, 4'-trichlorobiphenyl). The analysis of measurement results produced independently by different laboratories is often described as *meta-analysis* (Higgins et al., 2009; Rukhin, 2013).

The measurement model is an observation equation: a laboratory random effects model (Toman and Possolo, 2009, 2010), which represents the value of mass fraction measured by each laboratory as $w_j = \mu + \lambda_j + \varepsilon_j$ for $j = 1, \dots, n$, where $n = 6$ is the number of laboratories, μ denotes the measurand that is estimated by the consensus value, $\lambda_1, \dots, \lambda_n$ are the laboratory effects (assumed to be a sample from a Gaussian distribution with mean 0 and standard deviation τ), and $\varepsilon_1, \dots, \varepsilon_n$ represent measurement errors (also assumed to be outcomes of Gaussian random variables with mean 0 and standard deviations $\sigma_1, \dots, \sigma_n$).

The data are the measured values $\{w_j\}$, the associated standard uncertainties $\{u_j\}$, and the numbers of degrees of freedom $\{\nu_j\}$ that these standard uncertainties are based on. If the data were only the $\{w_j\}$ it would not be possible to distinguish the laboratory effects $\{\lambda_j\}$ from the measurement errors $\{\varepsilon_j\}$. As it is, we know that the absolute values of the $\{\varepsilon_j\}$ are generally comparable to the $\{u_j\}$, and conclude that any "excess variance" the $\{w_j\}$ may exhibit is attributable to the $\{\lambda_j\}$, comparable to τ in absolute value.

DerSimonian-Laird Procedure. The standard deviation τ, of the laboratory effects, may be estimated in any one of several different ways. DerSimonian and Laird (1986) suggested the procedure most widely used in meta-analysis to fit this type of model: it is implemented in function `rma` defined in R package `metafor` (Viechtbauer, 2010).

The fact that the estimate of τ is about three times larger than the median of the $\{u_j\}$, indicates that there is a source of uncertainty that has not been recognized by the participating laboratories, hence is not captured in their stated uncertainties. Thompson and Ellison (2011) call the contribution from such unrecognized source *dark uncertainty*, and in this case it is very substantial. The random effects model provides the technical machinery necessary to recognize and propagate this contribution.

The corresponding estimate of the measurand is $\widehat{\mu} = 33.6\,\text{ng/g}$. The same function `rma` also evaluates $u(\mu)$ as 0.75 ng/g, and produces a 95 % coverage interval for μ ranging from 32.3 ng/g to 34.9 ng/g.

An alternative, possibly more refined uncertainty evaluation that recognizes the limited numbers of degrees of freedom that the $\{u_j\}$ are based on, employs the parametric statistical bootstrap (Efron and Tibshirani, 1993), and produces $u(\mu) = 0.74\,\text{ng/g}$, as well as an approximate 95 % coverage interval ranging from 32.1 ng/g to 35.1 ng/g.

Bayesian Procedure. Both evaluations of uncertainty just discussed are over-optimistic because implicitly they regard an evaluation of the inter-laboratory variability τ that is based on five degrees of freedom only (since six laboratories are involved) as if it were based on infinitely many.

A Bayesian treatment can remedy this defect and recognize and propagate this source of uncertainty properly. The distinctive traits of a Bayesian treatment are these: (i) all quantities whose values are unknown are modeled as non-observable random variables, and data are modeled as observed values of random variables; (ii) estimates and uncertainty evaluations for unknown quantity values are derived from the conditional probability distribution of the unknowns given the data (the so-called *posterior distribution*).

Enacting (i) involves specifying probability distributions for all the quantities in play (unknowns as well as data), and (ii) involves application of Bayes's rule, typically via Markov Chain Monte Carlo sampling that produces an arbitrarily large sample from the posterior distribution (Gelman et al., 2013). Carrying this out successfully requires familiarity with probability models and with their selection for the intended purpose, and also with suitable, specialized software for statistical computing. The results reported below were obtained using function `metrop` defined in R package `mcmc` (Geyer and Johnson, 2014).

The prior distributions selected for the Bayesian analysis were these: μ has an (improper) uniform distribution over the set of its possible values; τ and the $\{\sigma_j\}$ have half-Cauchy distributions, the former with scale 15, the latter with scale 10, as suggested by Gelman (2006); the $\{\lambda_j\}$ are Gaussian with mean 0 and standard deviation τ. The data are modeled as follows: the $\{w_j\}$ are Gaussian with mean $\{\mu+\lambda_j\}$ and variances $\{\sigma_j^2\}$; and the $\{\nu_j u_j^2/\sigma_j^2\}$ are chi-squared with $\{\nu_j\}$ degrees of freedom.

The estimate of the consensus value μ is the mean of the corresponding posterior distribution, 33.6 ng/g, and the standard deviation of the same distribution is $u(\mu) = 0.99$ ng/g, which is substantially larger than the over-optimistic evaluation given above. A corresponding 95 % probability interval ranges from 31.5 ng/g to 35.5 ng/g.

Degrees of Equivalence. The Bayesian treatment greatly facilitates the characterization of the unilateral *degrees of equivalence* (DoE), which comprise the estimates of the $\{\lambda_j\}$ and the associated uncertainties $\{u(\lambda_j)\}$, depicted in Exhibit 14.

LAB	w_j/(ng/g)	u_j/(ng/g)	ν_j	LAB	w_j/(ng/g)	u_j/(ng/g)	ν_j
IRMM	34.30	1.03	60.0	NIST	32.42	0.29	2.0
KRISS	32.90	0.69	4.0	NMIJ	31.90	0.40	13.0
NARL	34.53	0.83	18.0	NRC	35.80	0.38	60.0

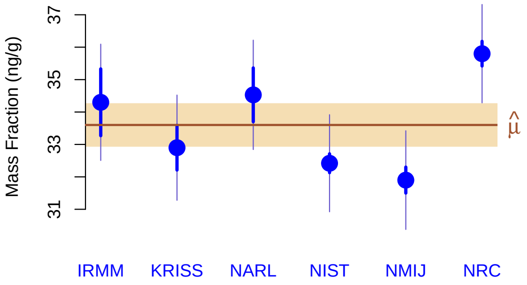


Exhibit 13: Measured values w_j of the mass fraction (ng/g) of PCB 28 in the sample, standard uncertainties u_j, and numbers of degrees of freedom ν_j that these standard uncertainties are based on, in CCQM–K25. Each large (blue) dot represents the value w_j measured by a participating laboratory; the thick, vertical line segment depicts $w_j \pm u_j$; and the thin, vertical line segment depicts the corresponding uncertainty including the contribution from *dark uncertainty*, $w_j \pm \left(\tau^2+u_j^2\right)^{½}$. The thick, horizontal (brown) line marks the consensus value $\hat{\mu}$, and the shaded (light-brown) band around it represents $\hat{\mu} \pm u(\mu)$.

E11 Microwave Step Attenuator. When a microwave signal is sent from a source to a load and their impedances are mismatched, some power is lost owing to reflections of the signal (Agilent, 2011). An attenuator (Exhibit 15) may then be used for impedance matching. Consider the following measurement model for the attenuation applied by a microwave

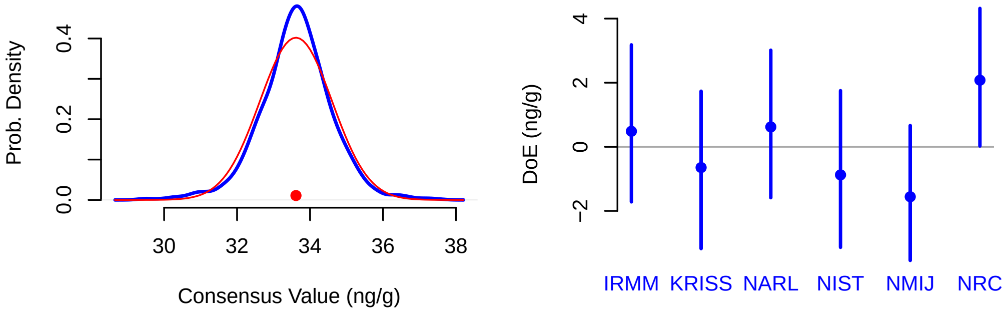


Exhibit 14: Bayesian posterior probability density of the consensus value (left panel), and unilateral degrees of equivalence (right panel). The red dot in the left panel marks the estimate of the consensus value, and the thin, bell-shaped, red curve is a Gaussian probability density with the same mean and standard deviation as the posterior density, showing that the posterior distribution has markedly heavier tails than the Gaussian approximation. The vertical line segments in the right panel correspond to 95 % probability intervals for the values of the degrees of equivalence, whose estimates are indicated by the blue dots.

coaxial attenuator (EA Laboratory Committee, 2013, Example S7):

$$L_{\mathrm{X}} = L_{\mathrm{S}} + \delta L_{\mathrm{S}} + \delta L_{\mathrm{D}} + \delta L_{\mathrm{M}} + \delta L_{\mathrm{K}} + \delta L_{\mathrm{ib}} - \delta L_{\mathrm{ia}} + \delta L_{\mathrm{0b}} - \delta L_{\mathrm{0a}}.$$

Exhibit 15: Coaxial step attenuator from Fairview Microwave Inc. (Allen, Texas) model SA3730N, performs attenuation from 0 dB to 30 dB in 1 dB steps for signals with frequency up to 3 GHz. The device is about 12 cm long and 7 cm tall. The large black knob controls attenuation in 10 dB steps, and the small black knob controls it in 1 dB steps.

The measurement serves to calibrate a microwave step attenuator using an attenuation measuring system containing a calibrated step attenuator which acts as the attenuation reference, and an analog null detector that is used to indicate the balance condition. The measurement method involves the determination of the attenuation between matched source and matched load. In this case the attenuator to be calibrated can be switched between nominal settings of 0 dB and 30 dB and it is this *incremental loss* that is determined in the calibration process (EA Laboratory Committee, 2013, S7.1).

The output is the attenuation L_X of the attenuator to be calibrated, and the inputs are as follows (with estimates and uncertainties listed in Exhibit 16):

- $L_{\mathrm{S}} = L_{\mathrm{ib}} - L_{\mathrm{ia}}$: difference in attenuation with the attenuator to be calibrated set at 30 dB (L_{ib}) and at 0 dB (L_{ia});
- δL_{S}: correction obtained from the calibration of the reference attenuator;

- δL_{D}: change in the attenuation of the reference attenuator since its last calibration due to drift;
- δL_{M}: correction due to mismatch loss;
- δL_{K}: correction for signal leakage between input and output of the attenuator to be calibrated, due to imperfect isolation;
- δL_{ia}, δL_{ib}: corrections to account for the limited resolution of the reference detector at the 0 dB and the 30 dB settings;
- δL_{0a}, δL_{0b}: corrections to account for the limited resolution of the null detector at the 0 dB and at the 30 dB settings.

L_{S} is estimated by the average, 30.0402 dB, of four observations of the attenuation difference aforementioned: 30.033 dB, 30.058 dB, 30.018 dB, and 30.052 dB. The standard uncertainty associated with that average is 0.0091 dB, given by the standard deviation of those four observations divided by $\sqrt{4}$, in accordance with Equation (A-5) of TN 1297.

EA Laboratory Committee (2013, Example S7) does not explain whether these repeated measurements were made after disconnecting the attenuators and then reconnecting them, or not. This is an important omission because connector repeatability actually is the dominant error in most microwave measurements.

Assuming that the perturbations expressed in the dispersion of these replicated readings are small, and considering that they are expressed in a logarithmic scale (decibel), we will proceed to model these four replicates as a sample from a Gaussian distribution. In these circumstances L_{S} is modeled as a Student t_3 random variable shifted to have mean 30.0402 dB and rescaled to have standard deviation 0.0091 dB.

Harris and Warner (1981) have shown that, under certain conditions, a U-shaped (arcsine) distribution is a suitable model for mismatch uncertainty, which derives from incomplete knowledge of the phase of the reflection coefficients of the source and load impedances, and of their interconnection. In this conformity, δL_{M} is modeled as a beta random variable with both parameters equal to ½, shifted to have mean 0, and rescaled to have standard deviation 0.02 dB. Exhibit 16 shows that the corresponding source of uncertainty makes a large contribution to measurement uncertainty, which is typical for microwave power transfers (Lymer, 2008).

δL_{D} could be comparably well modeled using either a Gaussian or an arcsine distribution, the latter being the more conservative choice that we have adopted.

δL_{K} is also modeled using a beta random variable with both parameters equal to ½, shifted to have mean 0, and rescaled to range between −0.003 dB and 0.003 dB. Such distribution has standard deviation 0.0021 dB (EA Laboratory Committee (2013, S7.12) lists 0.0017 dB instead, which is consistent with a rectangular distribution with the same range, but not with a U-shaped distribution that the same EA Laboratory Committee (2013, S7.12) indicates δL_{K} should have).

The Monte Carlo method of the GUM-S1 yields a distribution for the output quantity L_{X} that is markedly non-Gaussian (Exhibit 17), even though it is a linear combination of nine independent random variables: a situation that many users of the GUM would feel confident

QUANTITY	ESTIMATE (dB)	STD. UNC. (dB)	MODEL
L_S	30.0402	0.0091	Student t_3
δL_S	0.0030	0.0025	Rectangular
δL_D	0.0000	0.0014	Arcsine (U-shaped)
δL_M	0.0000	0.0200	Arcsine (U-shaped)
δL_K	0.0000	0.0021	Arcsine (U-shaped)
δL_ia	0.0000	0.0003	Rectangular
δL_ib	0.0000	0.0003	Rectangular
δL_0a	0.0000	0.0020	Gaussian
δL_0b	0.0000	0.0020	Gaussian

Exhibit 16: Estimates, standard uncertainties, and probability distributions for the input quantities that determine the value of the attenuation L_X of a microwave step attenuator. The arcsine distribution is a beta distribution with mean ½ and standard deviation $1/\sqrt{8}$, which here is re-scaled and shifted to reproduce the ranges and means of the corresponding random variables.

should give rise to a distribution close to Gaussian. The estimate of the output quantity is 30.043 dB, with associated standard uncertainty 0.0224 dB, which are the mean and standard deviation of a sample of size 1×10^7 drawn from the distribution of L_X (which reproduce the results listed in EA Laboratory Committee (2013, S7.12)).

A 95 % coverage interval for the true value of L_X ranges from 30.006 dB to 30.081 dB. The Monte Carlo sample may also be used to ascertain that the "conventional" 95 % coverage interval, of the form $L_\mathrm{X} \pm 2u(L_\mathrm{X})$, is conservative in this case, with effective coverage probability 99 %. However, the "conventional" 68 % coverage interval, of the form $L_\mathrm{X} \pm u(L_\mathrm{X})$, is too short, with effective coverage probability of only 61 %.

The bimodality of the distribution of the output quantity is attributable to the dominance of the contribution that the uncertainty associated with δL_M makes to the uncertainty associated with L_X: $u^2(\delta L_\mathrm{M})$ amounts to almost 79 % of $u^2(L_\mathrm{X})$. Not only is the distribution markedly non-Gaussian, but the shortest 68 % coverage "interval" turns out to be a union of two disjoint intervals, and does not even include the mean value of the distribution.

E12 Tin Standard Solution. A calibration standard intended to be used for the determination of tin was prepared gravimetrically by adding high-purity, assayed tin to an acidified aqueous solution, to achieve a mass fraction of tin of $w_G = 10.000\,07$ mg/g with standard uncertainty $u(w_G) = 0.010\,004\,77$ mg/g based on $\nu_G = 24$ degrees of freedom. The determination of the same mass fraction using inductively coupled plasma optical emission spectrometry (ICP-OES) yielded $w_I = 10.022\,39$ mg/g with standard uncertainty $u(w_I) = 0.010\,571\,82$ mg/g based on $\nu_I = 28$ degrees of freedom. In addition, the long-term (8 years) stability has been evaluated (Linsinger et al., 2001), and the associated uncertainty component has standard uncertainty $u_S(w) = 0.005\,823\,008$ mg/g based on $\nu_S = 55$ degrees of freedom.

Average. The most obvious combination of the two measurement results consists of averaging the measured values to obtain $a = ½(w_G + w_I) = 10.011$ mg/g and computing $u_A(a) = ½\big(u^2(w_G) + u^2(w_I)\big)^{½} = 0.007$ mg/g. Using a coverage factor $k = 2$ leads to an approximate 95 % coverage interval ranging from 9.997 mg/g to 10.026 mg/g.

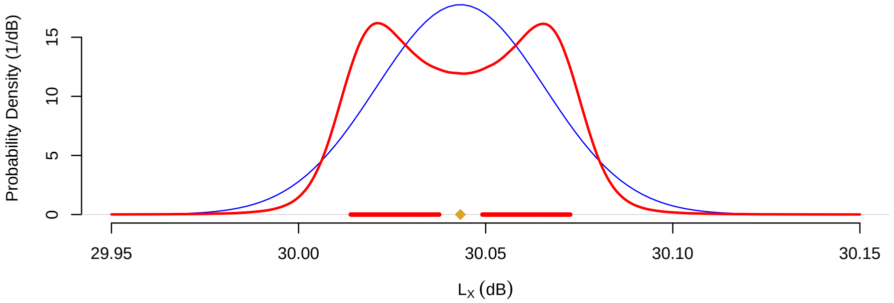


Exhibit 17: The curve with two humps is the probability density of the attenuation L_X applied by a microwave coaxial attenuator. The diamond marks the mean value of the distribution: it is not included in the union of the two intervals indicated by the thick horizontal line segments, which together define the shortest 68 % coverage region for the true value of that attenuation. The bell-shaped dotted curve is the probability density of a Gaussian distribution with the same mean and standard deviation as L_X.

Weighted Average. Alternatively, the weighted average with weights inversely proportional to the squared standard uncertainties is

$$a_W = \frac{\dfrac{w_G}{u^2(w_G)} + \dfrac{w_I}{u^2(w_I)}}{\dfrac{1}{u^2(w_G)} + \dfrac{1}{u^2(w_I)}} = 10.011\,\text{mg/g}, \text{ and } u_W(a) = \sqrt{\frac{1}{\dfrac{1}{u^2(w_G)} + \dfrac{1}{u^2(w_I)}}} = 0.011\,\text{mg/g}.$$

An approximate 95 % coverage interval $a_W \pm 2u_W(a)$ ranges from 9.996 mg/g to 10.025 mg/g.

GUM. Both $u_A(a)$ and $u_W(a)$ ignore the fact that the standard measurement uncertainties for the two measurement methods are based on small numbers of degrees of freedom. If the choice is made to combine the two measured values according to the measurement equation $a = ½(w_G + w_I)$ as considered above, then the GUM G.6.4 suggests that the coverage factor k, for a 95 % coverage interval of the form $a \pm k u_W(a)$, should take into account the numbers of degrees of freedom that $u(w_G)$ and $u(w_I)$ are based on.

This is accomplished by selecting k to be the 97.5th percentile of the Student's t distribution with

$$\nu^* = \frac{\left(u^2(w_G) + u^2(w_I)\right)^2}{\dfrac{u^4(w_G)}{\nu_G} + \dfrac{u^4(w_I)}{\nu_I}} = 51.76$$

degrees of freedom, according to the Welch-Satterthwaite formula (Miller, 1986, Page 61). The corresponding interval ranges from 9.997 mg/g to 10.026 mg/g.

Monte Carlo Method. To employ the Monte Carlo method of the GUM-S1, note that if z_G denotes a value drawn from a Student's t distribution with ν_G degrees of freedom, then $x = w_G + u(w_G) z_G \sqrt{(\nu_G - 2)/\nu_G}$ is like a drawing from the probability distribution of w_G. By the same token, $y = w_I + u(w_I) z_I \sqrt{(\nu_I - 2)/\nu_I}$ is like a drawing from the probability distribution of w_I, where z_I denotes a value drawn from Student's t distribution with ν_I degrees of freedom.

Repeating both drawings a sufficiently large number K of times, we may then form K replicates of the output quantity as $a_{S,1} = ½(x_1 + y_1), \dots, a_{S,K} = ½(x_K + y_K)$, whose standard deviation is an evaluation of $u_S(a)$, and whose 2.5th and 97.5th percentiles are the end-points of a 95,% coverage interval. With $K = 10^7$, we obtained $u_S(a) = 0.007\,\text{mg/g}$ and a 95 % coverage interval ranging from 9.997 mg/g to 10.026 mg/g.

Consensus. Finally, we consider a method of data reduction that blends the gravimetric and ICP-OES measurement results into a consensus estimate of the mass fraction of tin taking into account the difference between the values measured by the two methods. This method is currently used to assign a value, and to characterize the associated uncertainty, for NIST Standard Reference Materials that are single element solutions intended for use in spectrometry.

This approach is widely used in meta-analysis in medicine, where we seek to combine information from multiple, independent studies of a particular therapy or surgical procedure (Hedges and Olkin, 1985), and more generally to combine information from multiple sources about the same measurand (Gaver et al., 1992).

The measurement model is a set of two observation equations: $w_G = \omega + \lambda_G + \varepsilon_G$ and $w_I = \omega + \lambda_I + \varepsilon_I$, where λ_G and λ_I denote method effects specific to the gravimetry and to ICP-OES, and ε_G and ε_I represent measurement errors.

This is the simplest version of the measurement model used in several examples in this *Simple Guide*: E10, E21, E23, and E12: a linear, Gaussian random effects model. The model is linear because the quantities on the right-hand side of the observation equations are added. The model is Gaussian because the method (random) effects λ_G and λ_I are modeled as values of independent Gaussian random variables with mean 0 and the same standard deviation τ, and the measurement errors ε_G and ε_I are modeled as values of independent Gaussian random variables with mean 0 and standard deviations equal to $u(w_G)$ and $u(w_I)$.

Application of the most widely used procedure to fit such random effects model (DerSimonian and Laird, 1986), as implemented in function `rma` defined in R package `metafor` (Viechtbauer, 2010), produces 10.011 mg/g as consensus estimate.

The corresponding uncertainty evaluation may be done using a conventional approximation implemented in that same R function `rma`, or the parametric statistical bootstrap (Efron and Tibshirani, 1993), which is a version of the Monte Carlo method of the GUM-S1. R function `rma` (including the adjustment suggested by Knapp and Hartung (2003)) produces $u_D(a) = 0.011\,\text{mg/g}$, and a 95 % coverage interval that ranges from 9.869 mg/g to 10.153 mg/g.

The Monte Carlo evaluation of the uncertainty associated with the DerSimonian-Laird estimate involved the following steps.

1. Model the state of knowledge about τ^2 as an outcome of a random variable with a lognormal distribution with mean equal to the estimate $\widehat{\tau}^2 = 0.000\,143\,161\,8(\text{mg/g})^2$ produced by R function `rma`, and with standard deviation set equal to the estimate of the standard error of $\widehat{\tau}^2$, $0.000\,352\,268\,2(\text{mg/g})^2$, computed by the same function as explained by Viechtbauer (2007).

2. Select a sufficiently large integer K and then repeat the following steps for $k = 1, \dots, K$:

   (a) Draw a value τ_k^2 from the lognormal probability distribution associated with τ^2;

(b) Draw a value $\upsilon^2_{G,k}$ from a chi-squared distribution with ν_G degrees of freedom, and compute $\sigma_{G,k} = \left(\nu_G u^2(w_G)/\upsilon^2_{G,k}\right)^{½}$;

(c) Draw a value $\upsilon^2_{I,k}$ from a chi-squared distribution with ν_I degrees of freedom, and compute $\sigma_{I,k} = \left(\nu_I u^2(w_I)/\upsilon^2_{I,k}\right)^{½}$;

(d) Draw $\lambda_{G,k}$ and $\lambda_{I,k}$ from a Gaussian distribution with mean 0 and standard deviation τ_k;

(e) Draw $\varepsilon_{G,k}$ from a Gaussian distribution with mean 0 and standard deviation $\sigma_{G,k}$;

(f) Draw $\varepsilon_{I,k}$ from a Gaussian distribution with mean 0 and standard deviation $\sigma_{I,k}$;

(g) Compute $w_{G,k} = \widehat{\omega} + \lambda_{G,k} + \varepsilon_{G,k}$, and $w_{I,k} = \widehat{\omega} + \lambda_{I,k} + \varepsilon_{I,k}$;

(h) Compute the DerSimonian-Laird estimate w^*_k of ω based on $(w_{G,k}, \sigma_{G,k})$ and $(w_{I,k}, \sigma_{I,k})$.

A Monte Carlo sample $\{w^*_k\}$ of size $K = 50\,000$ drawn from the distribution of the mass fraction of tin as just described had standard deviation $u_B(a) = 0.012\,825\,58$ mg/g. A 95 % coverage interval derived from the Monte Carlo sample ranges from 9.986 075 mg/g to 10.035 862 mg/g.

Exhibit 18 shows the measurement results and the probability density for the measurand obtained by application of the Monte Carlo method to the DerSimonian-Laird consensus procedure. Exhibit 19 summarizes the results from the several different approaches discussed above.

E13 Thermal Expansion Coefficient. The thermal expansion coefficient of a copper bar is given by the measurement equation $\alpha = (L_1 - L_0)/(L_0(T_1 - T_0))$, as a function of the lengths $L_0 = 1.4999$ m and $L_1 = 1.5021$ m that were measured at temperatures $T_0 = 288.15$ K and $T_1 = 373.10$ K. The corresponding standard uncertainties are $u(L_0) = 0.0001$ m, $u(L_1) = 0.0002$ m, $u(T_0) = 0.02$ K, and $u(T_1) = 0.05$ K.

Gaussian Inputs. In the absence of information about the provenance of these estimates and uncertainty evaluations, we assign Gaussian distributions to them, with means equal to the estimates, and standard deviations equal to the standard uncertainties, and apply the NUM. Gauss's formula and the Monte Carlo method yield the same estimate and standard uncertainty for the thermal expansion coefficient: $\widehat{\alpha} = 1.73 \times 10^{-5}\,\mathrm{K}^{-1}$, and $u(\alpha) = 0.18 \times 10^{-5}\,\mathrm{K}^{-1}$.

A 95 % coverage interval for α can be derived from the Monte Carlo sample drawn from the probability distribution of the measurand, by selecting the 2.5th and 97.5th percentiles of the sample (of size 1×10^7) as end-points: $(1.38 \times 10^{-5}\,\mathrm{K}^{-1}, 2.07 \times 10^{-5}\,\mathrm{K}^{-1})$. A 99 % coverage interval built similarly ranges from $1.27 \times 10^{-5}\,\mathrm{K}^{-1}$ to $2.18 \times 10^{-5}\,\mathrm{K}^{-1}$.

Student Inputs. Now suppose that the estimates of the lengths and of the temperatures each is an average of four observations made under conditions of repeatability (VIM 2.20), and the corresponding standard uncertainties are the standard errors of these averages (Type A evaluations obtained by application of Equation (A-5) of TN1297).

In these circumstances, it may be more appropriate to assign Student t distributions with 3 degrees of freedom to all the inputs, shifted and scaled to have means and standard deviations equal to the corresponding estimates and standard uncertainties. The reason is this: if $\overline{x}$ and s denote the average and standard deviation of a sample of size m drawn from a Gaussian

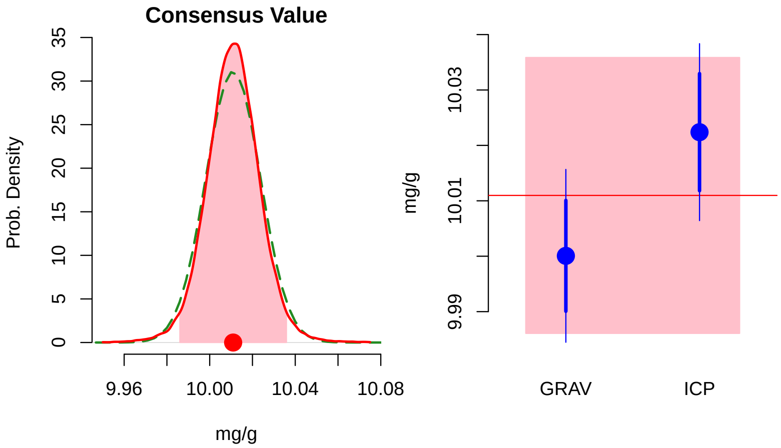


Exhibit 18: The left panel shows an estimate of the probability density of the consensus value, which fully characterizes measurement uncertainty. The consensus value $\widehat{\omega}$ is indicated by a large red dot. The pink area comprises 95 % of the area under the curve: its footprint on the horizontal axis is the corresponding coverage interval. The right panel shows the measurement results for the two methods used, gravimetry and ICP-OES, where the large red dots indicate the measured values, the thick, vertical blue lines indicate the corresponding standard measurement uncertainties, and the tiny thin lines that extend the thick lines indicate the contributions from *dark uncertainty* (between-methods uncertainty component (Thompson and Ellison, 2011)). The pink rectangle represents $\widehat{\omega} \pm U_{95\,\%}(\omega)$, where $U_{95\,\%}(\omega)$ denotes the expanded uncertainty corresponding to the specified coverage probability.

distribution with unknown mean μ and unknown standard deviation σ, then $(\bar{x}-\mu)/(s/\sqrt{m})$ has a Student's t distribution with $m-1$ degrees of freedom (DeGroot and Schervish, 2011, Theorem 8.4.2).

Both $\widehat{\alpha}$ and $u(\alpha)$ still have the same values as when the inputs are assigned Gaussian distributions, but the coverage intervals differ from those given above: the 95 % coverage interval constructed as described above ranges from $1.40 \times 10^{-5}\,\mathrm{K}^{-1}$ to $2.05 \times 10^{-5}\,\mathrm{K}^{-1}$, and the 99 % coverage interval ranges from $1.15 \times 10^{-5}\,\mathrm{K}^{-1}$ to $2.30 \times 10^{-5}\,\mathrm{K}^{-1}$.

Exhibit 20 shows the probability densities of the measurand that correspond to the two different modeling assumptions for the inputs. When these are modeled as Student's t_3 random variables, the distribution of the measurand is more concentrated around the mean, but also has heavier tails than when they are modeled as Gaussian random variables. This fact helps explain why the 95 % interval corresponding to the Student inputs is shorter than its counterpart for the Gaussian inputs, and that the opposite is true for the 99 % interval.

E14 Characteristic Strength of Alumina. Quinn and Quinn (2010) observed the following values of stress (expressed in MPa) when $m = 32$ specimens of alumina ruptured in a flexure test: 265, 272, 283, 309, 311, 320, 323, 324, 326, 334, 337, 351, 361, 366, 375, 380, 384, 389, 390, 390, 391, 392, 396, 396, 396, 396, 398, 403, 404, 429, 430, 435.

An adequate statistical model (observation equation) describes the data as outcomes of in-

APPROACH	ESTIMATE	STD. UNC.	95 % COV. INT.
AVE	10.011	0.007	(9.997, 10.026)
WAVE	10.011	0.007	(9.996, 10.025)
GUM	10.011	0.007	(9.997, 10.026)
GUM-S1	10.011	0.007	(9.997, 10.026)
DL	10.011	0.011	(9.869, 10.153)
DL-B	10.011	0.013	(9.986, 10.036)

Exhibit 19: Estimate of the mass fraction of tin in a standard solution based on two independent measurement results, obtained as a simple average (AVE), as a weighted average (WAVE), using the methods of the GUM and of the GUM-S1, and the consensus procedure (DL) suggested by DerSimonian and Laird (1986), as well as the same procedure but with uncertainty evaluation via the parametric statistical bootstrap (DL-B). All the values in the table are expressed in mg/g.

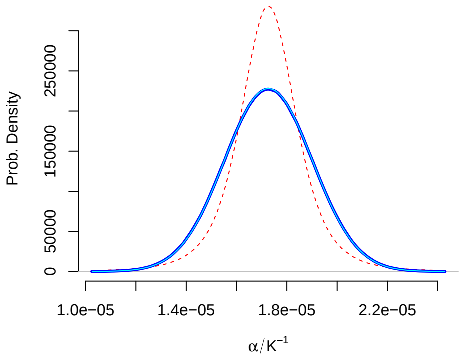


Exhibit 20: Estimates of the probability density of the thermal expansion coefficient assuming that the inputs have Student's t_3 distributions, shifted and rescaled to reproduce their estimated values and associated standard uncertainties (dashed, thin red line), or Gaussian distributions (solid, thick blue line). The solid, thin cyan line that essentially tracks the blue line, is the probability density of a Gaussian distribution with the same mean and standard deviation as the measurand.

dependent random variables with the same Weibull distribution with shape α and scale σ_C. A lognormal distribution would also be an acceptable model, but the Weibull is preferable according the Bayesian Information Criterion (BIC) (Burnham and Anderson, 2002).

The Weibull model may be characterized by saying that the rupture stress S of an alumina coupon is such that $\log S = \log \sigma_C + (1/\alpha) \log Z$, where Z denotes a measurement error assumed to have an exponential distribution with mean 1. Both the scale parameter σ_C and the shape parameter α need to be estimated from the data.

The measurand is σ_C, also called the *characteristic strength* of the material. Several different methods may be employed to estimate the shape and scale parameters. The maximum likelihood estimates are the values that maximize the logarithm of the likelihood function, which in this case takes the form

$$\ell(\alpha, \sigma_C) = m \log \alpha - m\alpha \log \sigma_C + (\alpha - 1) \sum_{i=1}^{m} \log s_i - \sum_{i=1}^{m} \left(\frac{x_i}{\sigma_C} \right)^{\alpha},$$

where $s_1, \ldots, s_m$ denote the rupture stresses listed above.

The maximum-likelihood estimates, determined by numerical optimization, are $\widehat{\alpha} = 10.1$ and $\widehat{\sigma}_{\text{C}} = 383$ MPa. The associated standard uncertainties are $u(\alpha) = 1.4$ and $u(\sigma_{\text{C}}) = 7.1$ MPa approximately. These approximations are derived from the curvature of the log-likelihood function ℓ at its maximum, according to the theory of the method (Wasserman, 2004). An approximate 95 % coverage interval for the characteristic strength ranges from 369 MPa to 398 MPa. The maximum likelihood estimates, the associated uncertainties, and this coverage interval, were computed using facilities of R package `bbmle` (Bolker and R Development Core Team, 2014).

Next consider a different measurand: the mean value η of the rupture stress. It is estimated as $\widehat{\eta} = \widehat{\sigma}_{\text{C}}\Gamma(1 + 1/\widehat{\alpha}) = 365$ MPa, where "Γ" denotes the gamma function (Askey and Roy, 2010). The equation $\eta = \sigma_{\text{C}}\Gamma(1+1/\alpha)$ is a measurement equation in its own right: since the maximum-likelihood calculation above provides approximations not only for the standard uncertainties associated with σ_{C} and with α, but also for their correlation coefficient (0.31), the NUM may then be used to find $u(\eta) \approx 8$ MPa.

The parametric statistical bootstrap (Monte Carlo method of the GUM-S1 and GUM-S2) (Efron and Tibshirani, 1993) may be used to evaluate the uncertainty associated with the pair $(\widehat{\alpha}, \widehat{\sigma}_{\text{C}})$ and with $\widehat{\eta}$. This is accomplished by first selecting a large number K of replicates to be generated for the quantities of interest (in this case, $K = 10\,000$). Next, for each $k = 1, \dots, K$, a sample of size $m = 32$ is drawn from a Weibull distribution with shape $\widehat{\alpha}$ and scale $\widehat{\sigma}_{\text{C}}$, and the corresponding maximum likelihood estimates α^*_k and $\sigma^*_{\text{C},k}$, and η^*_k, are computed in the same manner as for the original data. Finally, the resulting replicates are summarized as in Exhibit 21.

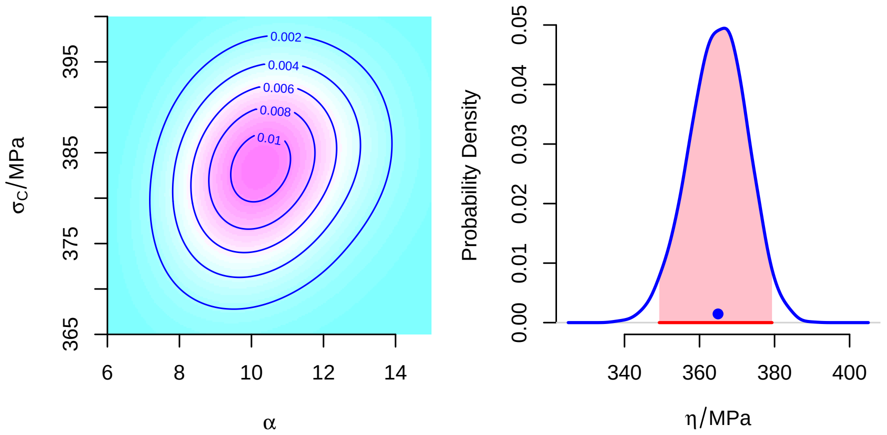


Exhibit 21: The left panel shows an estimate of the joint probability density of the maximum likelihood estimates of the scale and shape parameters of the Weibull distribution used to model the sampling variability of the alumina rupture stress. The right panel shows an estimate of the probability density of the mean rupture stress. The (pink) shaded area under the curve comprises 95 % of the total area under the curve, hence its projection onto the horizontal axis, marked by a thick, horizontal (red) line segment, is a 95 % coverage interval for η that ranges from 349 MPa to 379 MPa. The large (blue) dot marks the mean of the Monte Carlo sample, 365 MPa.

E15 Voltage Reflection Coefficient. Tsui et al. (2012) consider the voltage reflection coefficient $\Gamma = S_{22} - S_{12}S_{23}/S_{13}$ of a microwave power splitter, defined as a function of elements of the corresponding three-port scattering matrix (*S-parameters*). Exhibit 22 reproduces the measurement results for the S-parameters listed in Tsui et al. (2012, Table 5). Since the S-parameters (input quantities) are complex-valued, so is Γ (output quantity). Therefore, in this example the measurement model is a measurement equation with a vector-valued output quantity $(\Re(\Gamma), \Im(\Gamma))$ whose components are the real and imaginary parts of Γ.

The S-parameters are assumed to be independent, complex-valued random variables. The modulus and argument of each S-parameter are modeled as independent Gaussian random variables with mean and standard deviation equal to the value and standard uncertainty listed in Exhibit 22.

Application of the Monte Carlo method involves drawing samples of size K from the probability distributions of the four S-parameters, and using corresponding values from these samples to compute K replicates of Γ, which may then be summarized as in Exhibit 23 to characterize the associated uncertainty.

Since the real and imaginary parts of Γ both may be written as functions of the same eight input variables (which are the moduli and arguments of the S-parameters), the NUM may be used to generate "coupled" samples of the real and imaginary parts by treating them as elements of a vector-valued measurand.

It is also possible to incorporate correlations between the S-parameters, as well as correlations between the modulus and argument of any of the S-parameters, by specifying a suitable correlation matrix and applying it via one of the copulas (Possolo, 2010) that is available in the NUM.

Once the Monte Carlo samples produced by the NUM will have been saved, they may be imported into any statistical computing application to compute suitable summaries of the joint distribution of the real and imaginary parts of Γ, for example as depicted in Exhibit 23. The estimate of $\Re(\Gamma)$ is 0.0074 and $u(\Re(\Gamma)) = 0.0050$. The estimate of $\Im(\Gamma)$ is 0.0031 and $u(\Im(\Gamma)) = 0.0045$. The correlation between $\Re(\Gamma)$ and $\Im(\Gamma)$ is 0.0323.

	Mod(S)	u(Mod(S))	Arg(S)	u(Arg(S))
S_{22}	0.24776	0.00337	4.88683	0.01392
S_{12}	0.49935	0.00340	4.78595	0.00835
S_{23}	0.24971	0.00170	4.85989	0.00842
S_{13}	0.49952	0.00340	4.79054	0.00835

Exhibit 22: S-parameters expressed in polar form, and associated standard uncertainties, with Arg(S) and u(Arg(S)) expressed in radians.

E16 Oxygen Isotopes. Exhibit 24 reproduces the values of δ^{17}O and of δ^{18}O listed in Rumble et al. (2013, Table 2), which were determined in 24 samples of some of the oldest rocks on earth, part of the Isua Greenstone Belt near Nuuk, in southwestern Greenland.

Delta values (Coplen, 2011) express relative differences of isotope ratios in a sample and in a reference material, which for oxygen is the Vienna Standard Mean Ocean Water (VSMOW)

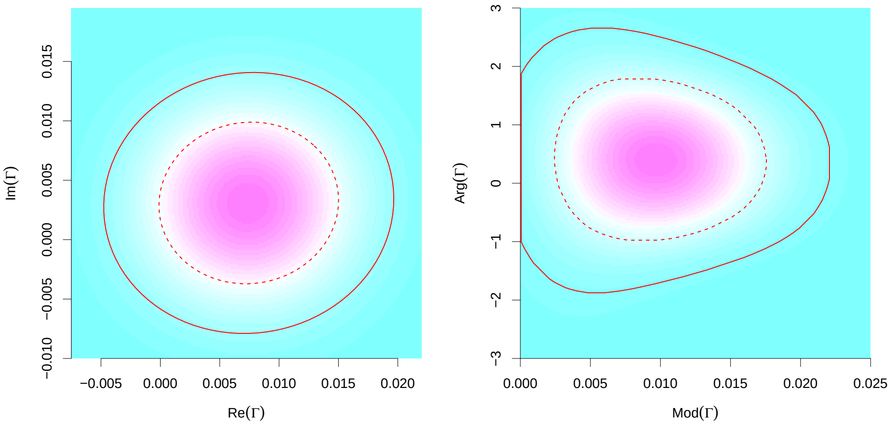


Exhibit 23: The left panel shows an estimate of the probability density of the joint distribution of the real and imaginary parts of Γ, and the right panel shows its counterpart for the modulus and argument of Γ. The solid curves outline 95 % (2σ) coverage regions, and the dashed curves outline 68 % (1σ) coverage regions.

δ^{17}O / ‰	5.24	6.02	3.92	4.29	5.66	7.32	2.52	5.34	2.57	6.11	1.23	0.97
δ^{18}O / ‰	10.09	11.56	7.54	8.31	10.86	14.11	4.92	10.30	5.01	11.77	2.37	2.02
δ^{17}O / ‰	1.10	5.23	1.45	3.42	2.85	3.32	7.13	5.17	6.87	5.65	6.57	2.50
δ^{18}O / ‰	2.19	10.08	2.74	6.58	5.49	6.40	13.67	9.96	13.19	10.83	12.49	4.75

Exhibit 24: Paired determinations of δ^{17}O and of δ^{18}O (expressed per mille) made on samples of rocks from the Isua Greenstone Belt (Rumble et al., 2013, Table 2).

maintained by the International Atomic Energy Agency (Martin and Gröning, 2009). For example, $\delta^{17}\text{O} = (R(^{17}\text{O}/^{16}\text{O})_\text{S} - R(^{17}\text{O}/^{16}\text{O})_\text{VSMOW})/R(^{17}\text{O}/^{16}\text{O})_\text{VSMOW}$, where $R(^{17}\text{O}/^{16}\text{O})_\text{S}$ denotes the ratio of the numbers of atoms of ^{17}O and of ^{16}O in a sample, and $R(^{17}\text{O}/^{16}\text{O})_\text{VSMOW} = 379.9 \times 10^{-6}$ (Wise and Watters, 2005a) is its counterpart for VSMOW.

Meijer and Li (1998) consider the following model for the relationship between δ^{17}O and δ^{18}O: $\log(1+\delta^{17}\text{O}) = \log(1+\kappa) + \lambda \log(1+\delta^{18}\text{O})$, where κ expresses the effect of imperfect calibration of the δ^{17}O scale to VSMOW (Meijer and Li, 1998, Page 362).

Exhibit 25 depicts the data listed in Exhibit 24, and a straight line fitted to the data by Deming regression, which is an errors-in-variables (EIV) model that recognizes that both sets of delta values have non-negligible and comparable measurement uncertainty (Adcock, 1878; Deming, 1943; Miller, 1981). The model was fitted to the data using function `mcreg` defined in R package `mcr` (Manuilova et al., 2014). The corresponding observation equations are as follows, for $i = 1, \dots, 24$:

$$\begin{aligned}\log(1 + \Delta_{17,i}) &= \log(1 + \kappa) + \lambda \log(1 + \Delta_{18,i}),\\ \delta^{17}\mathrm{O}_i &= \Delta_{17,i} + \varepsilon_{17,i},\\ \delta^{18}\mathrm{O}_i &= \Delta_{18,i} + \varepsilon_{18,i},\end{aligned}$$

where $\Delta_{17,i}$ and $\Delta_{18,i}$ denote the true values of the delta values for sample i, and $\varepsilon_{17,i}$ and $\varepsilon_{18,i}$ denote the corresponding measurement errors. These errors are modeled as Gaussian random variables with mean 0 and the same (unknown) standard deviation.

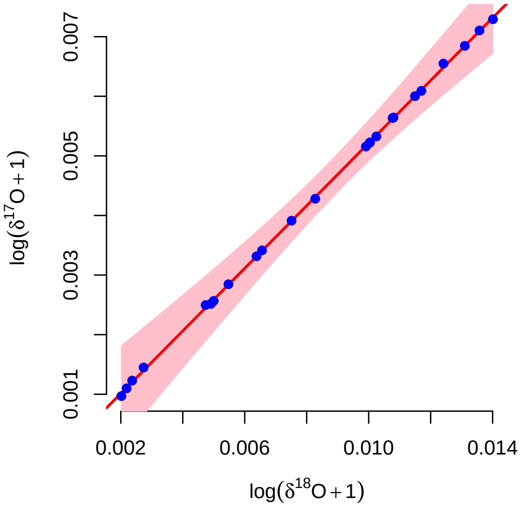


Exhibit 25: Straight line fitted by Deming regression to the paired determinations of δ^{17}O and of δ^{18}O made on samples of rocks from the Isua Greenstone Belt (Rumble et al., 2013, Table 2). The pink band is a 95 % simultaneous coverage band for the line: its thickness (vertical cross-section) has been magnified 25-fold.

For the data in Exhibit 24, the slope of the Deming regression line is $\widehat{\lambda} = 0.5253$. The theory of ideal, equilibrium mass-dependent fractionation suggests that $\lambda = 0.53$ (Matsuhisa et al., 1978; Weston Jr., 1999; Young et al., 2002).

The slope of the ordinary least squares (OLS) fit shares the same first four significant digits with the Deming regression slope in this case, because the data line up very closely to a straight line to begin with. In general, when there are errors in both variables, OLS produces an estimate of the slope whose absolute value is smaller than it should be — the so-called *regression attenuation* effect of ignoring the measurement uncertainty of the predictor, which is δ^{18}O in this case (Carroll et al., 2006).

The standard uncertainty associated with the slope of the Deming regression, based on a bootstrap sample of size $K = 100\,000$ was 0.0018. A 95 % coverage interval for its true value ranges from 0.5219 to 0.5289: these endpoints are the 2.5th and 97.5th percentiles of the bootstrap sample of the slope.

The uncertainty associated with the slope λ was evaluated by application of the non-parametric statistical bootstrap (Efron and Tibshirani, 1993), by repeating the following steps for $k = 1, \dots, K$, where $m = 24$ denotes the number of determinations of paired delta values:

1. Draw a sample of size m, uniformly at random and with replacement from the data $\{(\delta^{17}\mathrm{O}_i, \delta^{18}\mathrm{O}_i) : i = 1, \dots, m\}$, to obtain $\{(\delta^{17}\mathrm{O}_{i,k}, \delta^{18}\mathrm{O}_{i,k}) : i = 1, \dots, m\}$ — meaning that the m pairs of measured delta values are equally likely to go into the sample, and that any one of them may go into the sample more than once;

2. Fit a Deming regression line to the $\{(\delta^{17}\mathrm{O}_{i,k}, \delta^{18}\mathrm{O}_{i,k}) : i = 1, \dots, m\}$, and obtain its slope λ_k^*.

The standard uncertainty associated with λ is the standard deviation of $\lambda_1^*, \dots, \lambda_K^*$.

The estimate that Meijer and Li (1998) derived for λ, from measurements they made on a collection of samples of natural waters, was 0.5281, with standard uncertainty 0.0015. Since $(0.5253 - 0.5281)/\sqrt{0.0018^2 + 0.0015^2} = -1.2$, and the probability is 23 % that a Gaussian random variable with mean 0 and standard deviation 1 will deviate from 0 by this much or more to either side of 0, we conclude that the estimate of λ derived from the data in Exhibit 24 is statistically indistinguishable from the estimate obtained by Meijer and Li (1998).

Considering that one of these estimates is derived from the rocks of the Isua Greenstone Belt that are at least 4 billion years old, and that the other was derived from a collection of contemporary natural waters, Rumble et al. (2013) suggest that "the homogenization of oxygen isotopes required to produce such long-lived consistency was most easily established by mixing in a terrestrial magma ocean." It should be noted, however, that the exponent λ has been found to vary, albeit slightly, among various isotope fractionation processes (Barkan and Luz, 2011).

On the one hand, for the datasets considered by Meijer and Li (1998) and by Rumble et al. (2013), the estimated values of the constant κ in $\log(1+\delta^{17}\mathrm{O}) = \log(1+\kappa) + \lambda \log(1+\delta^{18}\mathrm{O})$ are very close to zero: they are 1.8×10^{-5} and -3.8×10^{-5}, respectively. On the other hand, typical values of $\delta^{18}\mathrm{O}$ (relative to VSMOW) range from -0.07 to 0.11.

These facts imply that the simplified relation $\delta^{17}\mathrm{O} = (1+\delta^{18}\mathrm{O})^{\lambda} - 1$ may be used to estimate the value of $\delta^{17}\mathrm{O}$ that corresponds to a given value of $\delta^{18}\mathrm{O}$, when it is reasonable to assume that there is equilibrium mass-dependent fractionation. Such estimate may be called for when computing the atomic weight of oxygen in a material for which only the value of $\delta^{18}\mathrm{O}$ has been measured, or when correcting measured values of $\delta^{13}\mathrm{C}$ in CO_2 for the $^{17}\mathrm{O}$ interference, when the measurements are made using an isotope-ratio mass spectrometer (Brand et al., 2010).

E17 Gas Analysis. Example 2 of ISO (2001, B.2.2) describes the estimation of an *analysis function* G that, given an instrumental response r as input, produces a value $x = G(r)$ of the amount-of-substance fraction of nitrogen in a synthetic version of natural gas. In this example, r denotes the indication produced by a thermal conductivity detector in a gas chromatograph.

The measurand is the analysis function G, which is determined based on the values of the amount fraction of nitrogen in a blank and in seven reference gas mixtures (standards), and on the corresponding instrumental responses, taken together with the associated uncertainties, all listed in Exhibit 26.

In ISO (2001, B.2.2), G is assumed to be a linear function that maps ρ_j (the true value of r_j) to $\xi_j = \alpha + \beta\rho_j$ (the true value of x_j), for $j = 1, \dots, n$, where $n = 8$ denotes the number of calibration data points. However, the uncertainty associated with the intercept α turns out to be about three times larger than the absolute value of the estimate of α, thus suggesting that the data are consistent with $\alpha = 0$.

A comparison of the two models for the analysis function, with and without intercept, via a formal analysis of variance performed disregarding the uncertainties associated with the instrumental responses (using R function `anova`), leads to the same conclusion: that the presence of α adds no value to the model. Therefore, in this example we use an analysis function G of the form $\xi_j = \beta\rho_j$.

The true values $\{\xi_j\}$ of the amount-of-substance fractions, and the true values $\{\rho_j\}$ of the corresponding instrumental responses, supposedly differ from their observed counterparts $\{x_j\}$ and $\{r_j\}$ that are listed in Exhibit 26, owing to measurement errors.

x	$u(x)$	r	$u(r)$	$\widehat{\xi}$	$\widehat{\rho}$
0.0015	0.00090	60	35.0	0.0015	61
0.1888	0.00045	7786	135.7	0.1888	7774
1.9900	0.00400	81700	36.7	1.9845	81711
3.7960	0.03900	156200	223.2	3.7937	156202
5.6770	0.01250	233300	137.2	5.6669	233329
7.1180	0.01250	293000	245.5	7.1165	293014
9.2100	0.02000	380600	125.1	9.2430	380569
10.9000	0.02500	449700	321.8	10.9200	449619

Exhibit 26: Amount-of-substance fraction x of nitrogen in a blank and in seven standards, corresponding instrumental responses r, and associated standard uncertainties $u(x)$ and $u(r)$, from ISO (2001, Table B.7), and estimates of their corresponding true values ξ (for x) and ρ (for r). The units for the amounts-of-substance fraction (x and $\widehat{\xi}$) and for $u(x)$ are µmol/mol. The instrumental responses (r and $\widehat{\rho}$) and $u(r)$ are dimensionless.

The measurement model for the analysis function comprises the following set of observation equations:

$$x_j = \xi_j + \varepsilon_j, \quad r_j = \rho_j + \delta_j, \quad \xi_j = \beta\rho_j, \quad \text{for } j = 1, \dots, n,$$

where the measurement errors $\varepsilon_1, \dots, \varepsilon_n$ and $\delta_1, \dots, \delta_n$ are assumed to be values of independent Gaussian random variables, all with mean 0, the former with standard deviations $u(x_1), \dots, u(x_n)$, and the latter with standard deviations $u(r_1), \dots, u(r_n)$.

ISO (2001) suggests that the values to be assigned to the unknown parameters (which are the slope β, and the true values $\rho_1, \dots, \rho_n$ of the instrumental responses) should be those that minimize

$$S(\beta, \rho_1, \dots, \rho_n) = \sum_{j=1}^{n} \left(\frac{x_j - \beta\rho_j}{u(x_j)}\right)^2 + \left(\frac{r_j - \rho_j}{u(r_j)}\right)^2.$$

This choice corresponds to maximum likelihood estimation, under the implied assumption that the standard uncertainties $\{u(x_j)\}$ and $\{u(r_j)\}$ are based on infinitely many degrees of freedom. Guenther and Possolo (2011) suggest an alternative criterion to be used when these numbers of degrees of freedom are finite and small, which often is the case.

The values of the arguments that minimize $S(\beta, \rho_1, \dots, \rho_n)$ are found by numerical optimization under the constraints that neither the true instrumental responses nor the true amount-of-substance fractions can be negative, using the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) as implemented in function `nloptr` defined in the R package of the same name (Johnson, 2015; Ypma, 2014).

The estimate of the slope is $\widehat{\beta} = 2.428\,724 \times 10^{-5}$ µmol/mol, and $S(\widehat{\beta}, \widehat{\rho}_1, \dots, \widehat{\rho}_n) = 6.16515$ (which is smaller than the corresponding value in ISO (2001, Page 26)). The estimates of the true values $\{\xi_j\}$ and $\{\rho_j\}$ are listed in the last two columns of Exhibit 26.

The uncertainty evaluation is done by application of the Monte Carlo method by repeating the following steps for $k = 1, \dots, K$ for a sufficiently large integer K:

1. Draw a sample value $x_{j,k}$ from a Gaussian distribution with mean $\widehat{\xi}_j$ and standard deviation $u(x_j)$, for $j = 1, \dots, n$.

2. Draw a sample value $r_{j,k}$ from a Gaussian distribution with mean $\widehat{\rho}_j$ and standard deviation $u(r_j)$, for $j = 1, \dots, n$.

3. Find the values β^*_k and $\rho^*_{1k}, \dots, \rho^*_{nk}$ that minimize $S(\beta, \rho_1, \dots, \rho_n)$ with $\{x_{j,k}\}$ and $\{r_{j,k}\}$ playing the roles of $\{x_j\}$ and $\{r_j\}$.

The standard deviation of the resulting $K = 10\,000$ replicates of the slope, $\beta^*_1, \dots, \beta^*_K$, was $u(\beta) = 2.3389 \times 10^{-8}$ µmol/mol, and a 95 % coverage interval for β ranges from 2.4242×10^{-5} to 2.4334×10^{-5}. The resulting probability densities of β and the $\{\rho_j\}$ are depicted in Exhibit 27.

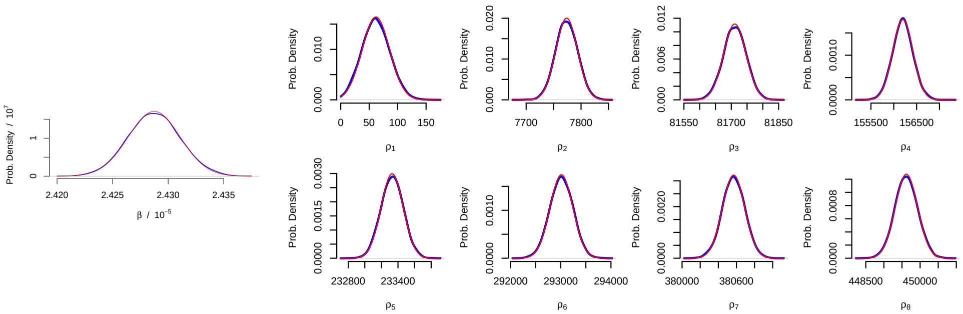


Exhibit 27: Estimates of the probability densities of β and of the $\{\rho_j\}$ (thick blue lines), and corresponding Gaussian probability densities with the same means and standard deviations (thin red lines).

ISO (2001, Table B.8) gives instrumental responses for two gas mixtures whose amount fractions of nitrogen are unknown. For $r_0 = 70\,000$, the analysis function estimated above produces $\widehat{\xi}_0 = 1.7001$ µmol/mol. The standard deviation of the K replicates $\{\beta^*_k r_0\}$ is $u(\xi_0) = 0.0016$ µmol/mol. For the second mixture, with $r_0 = 370\,000$, we obtain $\widehat{\xi}_0 = 8.9863$ µmol/mol and $u(\xi_0) = 0.0087$ µmol/mol. (Both standard uncertainties are smaller than their counterparts listed in the fourth column of ISO (2001, Table B.8).)

E18 Sulfur Dioxide in Nitrogen. NIST SRM 1693a Series M comprises ten 6 L (water volume) aluminum cylinders each containing about 0.85 m^3 (30 ft^3) at standard pressure and temperature, of a gas mixture with nominal amount fraction 50 µmol/mol of sulfur dioxide in nitrogen. The measuring instrument was a flow-through process analyzer with a pulsed UV fluorescence SO_2 detector (Thermo Scientific Model 43i-HL).

The measurement method used to assign values to the reference material in those ten cylinders involved (i) five primary standard gas mixtures (PSMs) with amount fractions of sulfur dioxide ranging from 40 µmol/mol to 60 µmol/mol, and (ii) a *lot standard*, which was a cylinder different from those ten, filled with the same gas mixture. Every time an instrumental indication was obtained either for a PSM or for one of the ten cylinders with the reference material, an indication was also obtained for the lot standard, and a ratio of the paired indications was computed.

Ten replicates of the ratio of instrumental indications were obtained for each of the five PSMs (Exhibit 28), and 21 replicates were obtained for each cylinder. Exhibit 30 shows boxplots of the ratios for all ten cylinders with the reference material, and also depicts the amount fractions and ratios for the standards, the associated uncertainties, and the analysis function that was built as described below. Exhibit 29 lists the values of the ratio for cylinder C05.

PSM	r	c	$u(c)$
S101	1.2089772	60.139	0.020
S101	1.2075386	60.139	0.020
S101	1.2030812	60.139	0.020
S101	1.2029003	60.139	0.020
S101	1.2045783	60.139	0.020
S101	1.2051815	60.139	0.020
S101	1.2099296	60.139	0.020
S101	1.2047862	60.139	0.020
S101	1.2072764	60.139	0.020
S101	1.2061604	60.139	0.020
S119	1.1051102	55.120	0.018
S119	1.1060079	55.120	0.018
S119	1.1028368	55.120	0.018
S119	1.1047117	55.120	0.018
S119	1.1085506	55.120	0.018
S119	1.1063356	55.120	0.018
S119	1.1040592	55.120	0.018
S119	1.1028799	55.120	0.018
S119	1.1025448	55.120	0.018
S119	1.1023260	55.120	0.018
S113	1.0050900	50.184	0.016
S113	1.0046148	50.184	0.016
S113	1.0039825	50.184	0.016
S113	1.0033240	50.184	0.016
S113	1.0058559	50.184	0.016
S113	1.0078670	50.184	0.016
S113	1.0049105	50.184	0.016
S113	1.0035410	50.184	0.016
S113	1.0067675	50.184	0.016
S113	1.0005778	50.184	0.016
S097	0.8968898	44.685	0.016
S097	0.8964397	44.685	0.016
S097	0.8958959	44.685	0.016
S097	0.8941153	44.685	0.016
S097	0.8924992	44.685	0.016
S097	0.8958868	44.685	0.016
S097	0.8933692	44.685	0.016
S097	0.8968200	44.685	0.016
S097	0.8950815	44.685	0.016
S097	0.8957939	44.685	0.016
S117	0.7958206	39.862	0.015
S117	0.7958588	39.862	0.015
S117	0.7953195	39.862	0.015
S117	0.7944788	39.862	0.015
S117	0.7939727	39.862	0.015
S117	0.7986906	39.862	0.015
S117	0.7981194	39.862	0.015
S117	0.7953721	39.862	0.015
S117	0.8006621	39.862	0.015
S117	0.7970862	39.862	0.015

Exhibit 28: For each replicate measurement of a PSM: the ratio r between the instrumental indications for the standard and for the lot standard, the amount fraction c of SO_2 in the standard, and the associated standard uncertainty $u(c)$.

r	DAY	r	DAY	r	DAY
0.9902957	1	0.9909091	2	0.9885518	3
0.9905144	1	0.9917382	2	0.9922330	3
0.9951852	1	0.9934188	2	0.9919730	4
0.9927100	1	0.9895879	3	0.9915597	4
0.9911975	1	0.9926954	3	0.9916772	4
0.9921821	2	0.9912304	3	0.9944420	4
0.9917024	2	0.9926593	3	0.9911971	4

Exhibit 29: For each replicate measurement of the reference material in cylinder C05: the ratio r between the instrumental indications for the reference material and for the lot standard, and the day when the measurement was made.

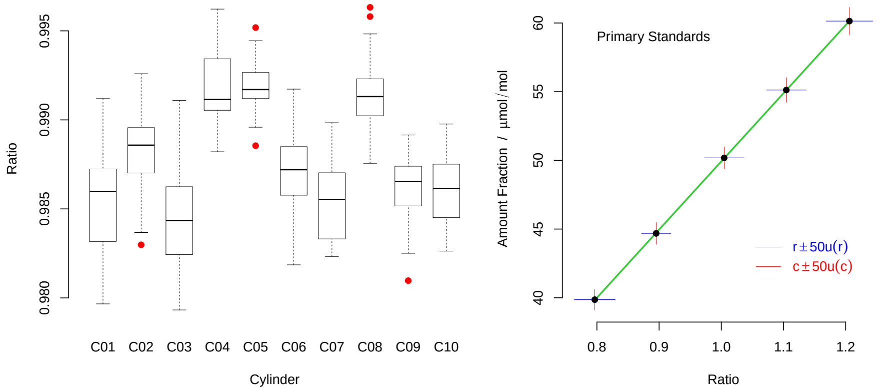


Exhibit 30: Boxplots of the ratios of instrumental readings for each cylinder (left panel), and amount fraction of SO_2 in the PSMs (right panel). Each boxplot summarizes 21 ratios determined for each cylinder under conditions of repeatability (VIM 2.20): the thick line across the middle of each box indicates the median, and the bottom and top of the box indicate the 25th and 75th percentiles of those ratios. Potential outlying ratios are indicated with large (red) dots. The horizontal and vertical line segments in the right panel depict the standard uncertainties associated with the ratios and with the amount fractions, magnified 50-fold. The sloping (green) line in the right panel is the graph of the analysis function.

The process used for value assignment comprised two steps: first, the data from the PSMs was used to build an analysis function that produced values of the amount fraction of sulfur dioxide given a value of the ratio aforementioned; second, this function was applied to the ratios pertaining to the 10 cylinders with the reference material.

Analysis Function. The procedure used to build the analysis function was suggested by Guenther and Possolo (2011): it is a modification of the procedure described in ISO 6143 (ISO, 2001), which is the internationally recognized standard used to certify gaseous reference materials. The procedure recognizes that the number of replicates of the ratios for the PSMs is modest (ten in this case), hence the Type A evaluations of the associated uncertainties, obtained by application of Equation (A-5) of TN1297, are based on only $10 - 1 = 9$ degrees of freedom.

Before the analysis function can be built, a functional form needs to be chosen for it: experience with these materials suggests that a polynomial of low degree affords an adequate model. The choice of degree for this polynomial was guided by diagnostic plots of the residuals corresponding to candidate models, supplemented by consideration of the *Bayesian Information Criterion* (BIC) (Burnham and Anderson, 2002). In this case, this amounts to selecting the polynomial of degree $p - 1$ for which $S(\beta_0, \beta_1, \rho_1, \dots, \rho_m) + (m + p) \log m$ is a minimum, where the function S is defined below.

The best model for the analysis function g turns out to be a first degree polynomial: $g(r) = \beta_0 + \beta_1 r$, which is depicted in the right panel of Exhibit 30. The coefficients β_0 (intercept) and β_1 (slope) were not estimated by ordinary least squares, but as the values that minimize the following criterion (Guenther and Possolo, 2011) that takes into account the fact that the standard uncertainties of the ratios are based on a finite number of degrees of freedom:

$$S(\beta_0, \beta_1, \rho_1, \dots, \rho_m) = \sum_{i=1}^{m} \left[\frac{\left(c_i - (\beta_0 + \beta_1 \rho_i)\right)^2}{2u^2(c_i)} + \frac{\nu_i + 1}{2} \log \left(1 + \frac{(r_i - \rho_i)^2}{\nu_i u^2(r_i)} \right) \right],$$

where $m = 5$ is the number of PSMs, $c_1, \dots, c_m$ are the amount fractions of SO_2 in the PSMs, $u(c_1), \dots, u(c_m)$ are the associated uncertainties, $r_1, \dots, r_m$ are the averages of the replicates of the ratios obtained for each PSM and $u(r_1), \dots, u(r_m)$ are the Type A evaluations of the associated uncertainties, $\nu_1, \dots, \nu_m$ are the numbers of degrees of freedom that the $\{u(r_i)\}$ are based on, and $\rho_1, \dots, \rho_m$ are the true values of the ratios.

The minimization procedure yields not only estimates of the intercept and slope of the analysis function, but also estimates of the true values of the ratios, $\rho_1, \dots, \rho_m$. This particular version of the more general criterion suggested by Guenther and Possolo (2011) is appropriate because the uncertainties associated with the amount fractions of SO_2 in the PSMs may be assumed to be based on large numbers of degrees of freedom, and only the uncertainties associated with the ratios are based on small numbers of degrees of freedom.

Uncertainty Evaluation. The uncertainty evaluation is performed by application of the Monte Carlo method as follows.

1. Choose a suitably large integer K (in this example $K = 5000$), and then for $k = 1, \dots, K$:

(a) Simulate $c_{1,k}$, …, $c_{m,k}$ as realized values of m independent Gaussian random variables with means equal to the amount fractions of SO_2 in the PSMs c_1, …, c_m, and standard deviations equal to the corresponding standard uncertainties $u(c_1), \dots, u(c_m)$, respectively.

(b) Simulate $w_{1,k}, \dots, w_{m,k}$ as realized values of of m independent chi-squared random variables with ν_1, …, ν_m degrees of freedom, respectively, and compute perturbed versions of the standard uncertainties associated with r_1, …, r_m as $u_k(r_1) = u(r_1)\sqrt{\nu_1/w_{1,k}}, \dots, u_k(r_m) = u(r_m)\sqrt{\nu_m/w_{m,k}}$.

(c) Simulate $r_{1,k}$, …, $r_{m,k}$ as realized values of m independent Gaussian random variables with means r_1, …, r_m, and standard deviations $u_k(r_1)$, …, $u_k(r_m)$, respectively.

(d) Minimize the criterion S defined above, with the $\{c_{i,k}\}$, $\{r_{i,k}\}$, and $\{u_k(r_i)\}$ in the roles of the $\{c_i\}$, $\{r_i\}$, and $\{u(r_i)\}$. Let $\beta^*_{0,k}$, $\beta^*_{1,k}$, $\rho^*_{1,k}$, …, $\rho^*_{m,k}$ denote the values at which the minimum is achieved.

Exhibit 31 depicts the probability densities of the Monte Carlo distributions of the intercept and slope of the analysis function. However, in the next step the Monte Carlo sample, $(\beta^*_{0,1}, \beta^*_{1,1}), \dots, (\beta^*_{0,K}, \beta^*_{1,K})$, will be used directly.

2. Suppose that r_1, …, r_n denote replicated determinations of the ratio for a particular cylinder with the reference material, for example, the 21 replicates listed in Exhibit 29 for cylinder C05. Since the boxplot depicting the ratios for this cylinder (Exhibit 30) suggests two potentially outlying values, these are set aside and not used in the next step, hence $n = 21 - 2 = 19$.

3. For each replicate $j = 1, \dots, n$ of the ratio for cylinder C05, compute K Monte Carlo replicates of the corresponding amount fraction of SO_2 as $c_{j,1} = \beta^*_{0,1} + \beta^*_{1,1}c_j$, …, $c_{j,K} = \beta^*_{0,K} + \beta^*_{1,K}c_j$.

4. Exhibit 31 shows the probability density of the $nK = 19 \times 5000 = 95\,000$ replicates of the amount fraction of SO_2 that correspond to the ratios in Exhibit 29. The standard deviation of this sample is an evaluation of $u(c) = 0.12\,\mu\text{mol/mol}$ for cylinder C05.

5. However, considering the use intended for this reference material, its long-term instability must also be recognized, and its effects incorporated in the uncertainty assessment. In this case, long-term instability is an important source of uncertainty, with corresponding standard uncertainty of 0.19 µmol/mol resulting from a Type B evaluation. It was propagated using the Monte Carlo method by adding Gaussian perturbations $z_{j,1}, \dots, z_{j,K}$ to the Monte Carlo sample of amount fractions, for each replicate $j = 1, \dots, n$ of the ratio of instrumental indications for cylinder C05, with mean 0 µmol/mol and standard deviation 0.19 µmol/mol, finally to obtain $x_{j,1} = c_{j,1} + z_{j,1}$, …, $x_{j,K} = c_{j,K} + z_{j,K}$.

6. The standard deviation of the $\{x_{j,k}\}$, 0.20 µmol/mol for this cylinder, was the evaluation of standard uncertainty. A 95 % coverage interval, built subject to the constraints that it be symmetrical around the estimated amount fraction $\widehat{c} = 49.52\,\mu\text{mol/mol}$, ranges from 49.13 µmol/mol to 49.91 µmol/mol. The corresponding expanded uncertainty is $U_{95\,\%}(c) = 0.40\,\mu\text{mol/mol}$, hence the effective (*post hoc*) coverage factor is $k = 2$.

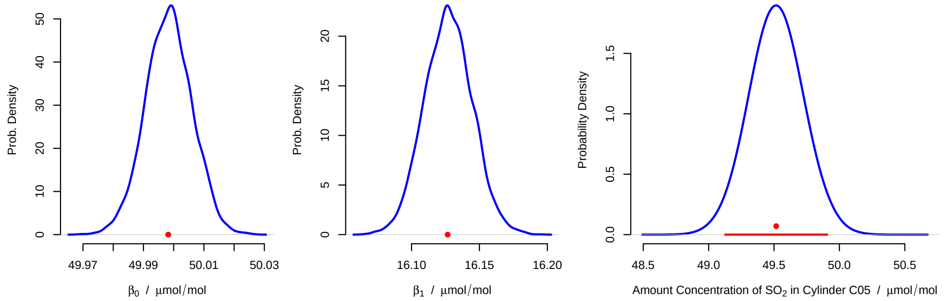


Exhibit 31: Probability density estimates for the intercept β_0 and slope β_1 of the analysis function, based on a Monte Carlo sample of size $K = 5000$ (left and center panels). Probability density estimate for the amount fraction of SO_2 in cylinder C05, taking into account uncertainty contributions from the calibration process and from long-term instability of the material. The value assigned to the reference material in this cylinder is indicated by a (red) dot, and the thick, horizontal (red) line segment represents a 95 % coverage interval (right panel).

E19 Thrombolysis. It is common medical practice to administer a drug that dissolves blood clots (that is, stimulates thrombolysis) to patients who suffer a myocardial infarction ("heart attack"). Exhibit 32 lists the results of a placebo-controlled randomized experiment that was carried out in Scotland between December 1988 and December 1991, to measure the decrease in the odds of death resulting from the administration of anistreplase (a thrombolytic agent) immediately at a patient's home rather than only upon arrival at the hospital (GREAT Group, 1992).

Group		HOME	HOSPITAL
Outcome	DEATH	13	23
	SURVIVAL	150	125

Exhibit 32: Results of an experiment to measure the efficacy of early administration of anistreplase in response to myocardial infarction. Patients were assigned at random, as if by tossing a fair coin, to the HOME and HOSPITAL groups. The 163 patients in the HOME group received the drug at home immediately upon diagnosis and the placebo at the hospital, while the opposite was done for the 148 patients in the HOSPITAL group. The recorded deaths are those that occurred within three months of the infarction.

The naive estimate of the probability of death for the early treatment (HOME group) is $p_E = 13/163$, and the corresponding *odds* of death are $o_E = p_E/(1-p_E) = 13/150$. Similarly, for the late treatment (HOSPITAL group), $o_L = 23/125$. The odds-ratio is $\text{OR} = o_E/o_L = 0.471$, and the *log-odds ratio* (which is the measurand) is $\theta = \log(\text{OR}) = -0.753$. The fact that $\text{OR} < 1$ (or, $\theta < 0$) suggests that the early treatment reduces the odds of death.

The sampling distribution of the log-odds ratio is approximately Gaussian with variance $u^2(\theta) \approx 1/13 + 1/23 + 1/150 + 1/125$, hence $u(\theta) \approx 0.37$ (Jewell, 2004, 7.1.2). Since

$\theta/u(\theta) = -2.05$, and the probability is 2 % that a Gaussian random variable with mean 0 and variance 1 will take a value less than -2.05, the conventional conclusion is that the early treatment reduces the odds of death significantly and substantially (by about $e^{\theta} - 1 = 53\,\%$) relative to the late treatment.

Since this is a surprisingly large reduction in the odds of death, Pocock and Spiegelhalter (1992) conjecture that "perhaps the Grampian region anistreplase trial was just lucky". The trial was terminated early, before the number of patients deemed necessary were recruited, possibly when the results were particularly favorable and seemed to have established the improvement incontrovertibly. Maybe early administration of anistreplase might not have appeared to have caused as striking an improvement if the study had been allowed to run its course.

Driven by this skepticism, Pocock and Spiegelhalter (1992) and Spiegelhalter et al. (2004, Example 3.6) describe an alternative Bayesian analysis that takes into account the belief of a senior cardiologist that a 15 % to 20 % reduction in mortality is highly plausible, while no benefit or detriment, as well as a relative reduction of more than 40 %, both are rather unlikely.

If this belief is described by a Gaussian distribution for the log-odds ratio whose central 95 % probability lies between $\log(1 - 0.4) = -0.51$ and 0, then the corresponding prior distribution for θ has mean $\theta_0 = -0.255$ and standard deviation $u(\theta_0) = 0.13$.

Applying Bayes's rule with this prior distribution and with the likelihood function corresponding to the aforementioned Gaussian sampling distribution for the log-odds ratio, leads to a Gaussian posterior distribution for θ (Spiegelhalter et al., 2004, Equation (3.15)) with mean

$$\frac{\dfrac{\theta}{u^2(\theta)} + \dfrac{\theta_0}{u^2(\theta_0)}}{\dfrac{1}{u^2(\theta)} + \dfrac{1}{u^2(\theta_0)}} = -0.31,$$

which suggests a reduction in the odds of death by only $e^{-0.311} - 1 = 27\,\%$.

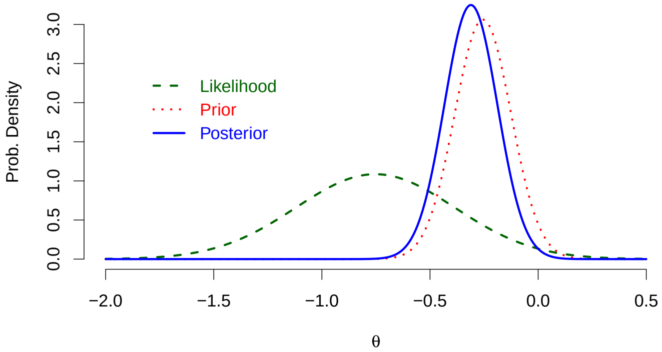


Exhibit 33: Likelihood function, prior and posterior probability densities for the measurement of the relative effect of early versus late administration of a thrombolytic agent.

Since the posterior distribution has standard deviation $[1/(1/u^2(\theta) + 1/u^2(\theta_0))]^{½} = 0.123$, a 95 % Bayesian coverage interval (*credible interval*) for the reduction in the odds of death ranges from 7 % to 42 %. Exhibit 33 depicts the prior density, the likelihood function, and the posterior density, making clear that the data nudge the informative prior distribution to the left (that is, toward smaller values of θ) only slightly. The spread of the posterior distribution is just a little smaller than the spread of the prior distribution, owing to the considerable uncertainty in the results given the fairly small total number (36) of deaths.

The Bayesian coverage interval may be compared with the conventional (sampling-theoretic) 95 % confidence interval for the odds ratio that is based on unconditional maximum likelihood estimation and Wald's Gaussian approximation, according to which the reduction in the odds of death ranges from 26 % to 163 %. This interval reflects both the aforementioned, possibly excessive optimism, and also the considerable uncertainty that again is attributable to the small total number of deaths that occurred. (This interval was computed using function `oddsratio` defined in R package `epitools` (Aragón, 2012).)

E20 Thermal Bath. The readings listed and depicted in Exhibit 34 were taken every minute with a thermocouple immersed in a thermal bath during a period of 100 min, to characterize the state of thermal equilibrium and to estimate the mean temperature of the bath.

The observation equation $t_i = \tau + \varepsilon_i + \varphi t_{i-1} + \theta_1\varepsilon_{i-1} + \theta_2\varepsilon_{i-2}$ models the sequence of observations as an auto-regressive, moving average (ARMA) time series, where the $\{\varepsilon_i\}$ are assumed to be independent and Gaussian with mean 0 and standard deviation σ.

Correlations between the $\{t_i\}$ arise because each reading of temperature depends on previous readings and on the errors that affect them. This particular ARMA model was selected according to Akaike's Information Criterion corrected for the finite length of the series (AICc) (Burnham and Anderson, 2002).

The maximum-likelihood estimates of the parameters, obtained using R function `arima`, are $\widehat{\tau} = 50.1054$ °C, $\widehat{\varphi} = 0.8574$, $\widehat{\theta}_1 = -0.5114$, $\widehat{\theta}_2 = 0.3369$, and $\widehat{\sigma} = 0.002$ °C. Furthermore, $u(\tau) = 0.001$ °C, which is about three times larger than the naive (and incorrect) evaluation that would have been obtained had the auto-correlations been neglected.

The results suggest that the variability of the bath's temperature includes a persistent pattern of oscillations, characterized by the auto-regressive parameter φ, possibly driven by imperfect insulation and convection. In addition, there are superimposed volatile effects, characterized by the moving average parameters θ_1 and θ_2, and by the "innovations" standard deviation σ.

E21 Newtonian Constant of Gravitation. Newton's law of universal gravitation states that two material objects attract each other with a force that is directly proportional to the product of their masses and inversely proportional to the squared distance between them. G is the constant of proportionality: it was first measured by Cavendish (1798).

Mohr et al. (2012) explain that because there is no known quantitative theoretical relationship between G and the other fundamental constants, a consensus estimate of G depends only on measurement results for G and not on measurement results for any of the other fundamental physical constants.

The value that the CODATA Task Group on Fundamental Constants recommends for G in the 2014 adjustment is $6.674\,08 \times 10^{-11}\ \mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$, with associated standard uncertainty $u(G) = 0.000\,31 \times 10^{-11}\ \mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$ (`physics.nist.gov/cuu/Constants`). The recommended value is a weighted average of the estimates of G listed in Exhibit 35, computed

0.1024	0.1054	0.1026	0.1042	0.1026	0.1039	0.1065	0.1052	0.1067	0.1072
0.1054	0.1049	0.1082	0.1039	0.1052	0.1085	0.1088	0.1075	0.1085	0.1098
0.1070	0.1060	0.1067	0.1065	0.1072	0.1062	0.1085	0.1062	0.1034	0.1049
0.1044	0.1057	0.1060	0.1082	0.1052	0.1060	0.1057	0.1072	0.1072	0.1077
0.1103	0.1090	0.1077	0.1082	0.1067	0.1098	0.1057	0.1060	0.1019	0.1021
0.0993	0.1014	0.0965	0.1014	0.0996	0.0993	0.1003	0.1006	0.1026	0.1014
0.1039	0.1044	0.1024	0.1037	0.1060	0.1024	0.1039	0.1070	0.1054	0.1065
0.1072	0.1065	0.1085	0.1080	0.1093	0.1090	0.1128	0.1080	0.1108	0.1085
0.1080	0.1100	0.1065	0.1062	0.1057	0.1052	0.1057	0.1034	0.1037	0.1009
0.1009	0.1044	0.1021	0.1021	0.1029	0.1037	0.1049	0.1082	0.1044	0.1067

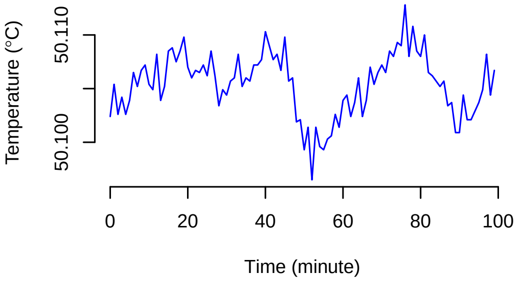


Exhibit 34: Time series of temperature readings (expressed as deviations from 50 °C, all positive) produced every minute by a thermocouple immersed in a thermal bath. The temporal order is from left to right in each row, and from top to bottom between rows. Data kindly shared by Victor Eduardo Herrera Diaz (*Centro de Metrología del Ejército Ecuatoriano*, CMME, Quito, Ecuador) during an international workshop held at the *Laboratorio Tecnológico del Uruguay* (LATU, Montevideo, Uruguay) in March, 2013.

similarly to the value recommended in the previous release (Mohr et al., 2012).

An alternative data reduction may be undertaken using methods that have been widely used to combine the results of multiple studies, in particular in medicine, where such combination is known as *meta-analysis* (Cooper et al., 2009). Usually, these studies are carried out independently of one another, hence pooling the results broadens the evidentiary basis for the conclusions at the same time as it reduces the associated uncertainty.

This alternative data reduction rests on an observation equation (statistical model) for the measurement results which, for laboratory or experiment j, comprises an estimate x_j of G and an evaluation u_j of the associated standard uncertainty, for $j = 1, \dots, n = 14$.

The statistical model expresses the value measured by laboratory j as $x_j = G + \lambda_j + \varepsilon_j$, where λ_j denotes an effect specific to the laboratory, and ε_j denotes measurement error. Both the laboratory effects $\lambda_1, \dots, \lambda_n$ and the measurement errors $e_1, \dots, e_n$ are modeled as outcomes of random variables with mean zero. The $\{\lambda_j\}$ all have the same standard deviation τ, but the $\{\varepsilon_j\}$ may have different standard deviations $\{u_j\}$.

This model achieves consistency between the measured values by adding unknown laboratory effects, $\{\lambda_j\}$, to the measured values. The approach adopted by CODATA achieves the same goal by applying a multiplicative factor (larger than 1) to the $\{u_j\}$, thus reflecting the belief that these standard uncertainties are too small given the dispersion of the measured values and assuming that all laboratories are measuring the same quantity.

Because the laboratory effects are modeled as random variables, the measurement model is

	G	$u(G)$
	$/1 \times 10^{-11}$ m^3 kg^{-1} s^{-2}	
NIST-82	6.672 482	0.000 428
TRD-96	6.672 900	0.000 500
LANL-97	6.673 984	0.000 695
UWash-00	6.674 255	0.000 092
BIPM-01	6.675 590	0.000 270
UWup-02	6.674 220	0.000 980
MSL-03	6.673 870	0.000 270
HUST-05	6.672 220	0.000 870
UZur-06	6.674 252	0.000 124
HUST-09	6.673 490	0.000 180
JILA-10	6.672 340	0.000 140
BIPM-13	6.675 540	0.000 160
ROSI-14	6.671 910	0.000 990
UCI-14	6.674 350	0.000 126

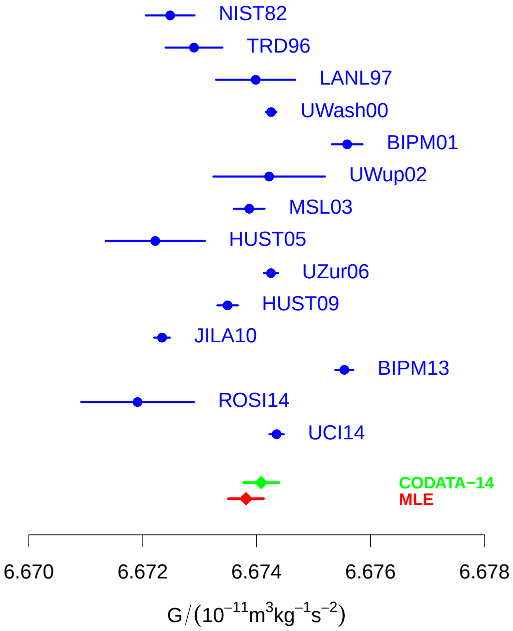


Exhibit 35: Values of G and $u(G)$ used to determine the 2014 CODATA recommended value (David Newell, 2015, Personal Communication) (left panel). The measurement results are depicted (right panel) in blue, with a dot indicating the measured value, and the horizontal line segment representing the interval $x_j \pm u_j$ where x_j denotes the value measured by experiment j and u_j denotes the associated uncertainty, for $j = 1, \dots, n = 14$. The 2014 CODATA recommended value and associated standard uncertainty, and their counterparts obtained via maximum likelihood estimation, are depicted similarly. Obviously, the 2014 CODATA recommended value and the maximum likelihood estimate are statistically indistinguishable.

called a *random effects model*. Mandel and Paule (1970), Mandel and Paule (1971), Rukhin and Vangel (1998), Toman and Possolo (2009), and Toman and Possolo (2010) discuss the use of models of this kind in measurement science, and Higgins et al. (2009) review them in general. The random variables $\{\lambda_j\}$ and $\{\varepsilon_j\}$ are usually assumed to be Gaussian and independent, but neither assumption is necessary.

When this model is used to estimate the value of G in the context of the CODATA adjustment, correlations between some of the laboratory effects need to be taken into account. In some applications, either the laboratory effects, or the measurement errors, or both, have non-Gaussian distributions (Pinheiro et al., 2001; Rukhin and Possolo, 2011).

The model may be fitted to the data listed in Exhibit 35 using any one of several different statistical procedures. For example, DerSimonian and Laird (1986) introduced one of the more widely used procedures, and Toman (2007) describes a Bayesian procedure. The more popular procedures assume that the laboratory effects and the errors are mutually independent. Since, in this case, the laboratory effects for NIST-82 and LANL-97 are correlated with correlation coefficient 0.351, and the laboratory effects for HUST-05 and HUST-07 are correlated with correlation coefficient 0.234 (Mohr et al., 2012, Pages 1568–1569), the more popular fitting procedures are not applicable here.

The method of maximum likelihood estimation may be used to fit the model to the data even in the presence of such correlations. (Bayesian methods, and variants of the DerSimonian-Laird procedure can do the same.) The general idea of maximum likelihood estimation is to choose values for the quantities whose values are unknown (G and τ in this case) that maximize the probability density of the data. Application of this method requires that the probability distribution of the random variables that the data are conceived as realized values of, be modeled explicitly.

We assume that the joint probability distribution of the $\{x_j\}$ is multivariate Gaussian with n-dimensional mean vector all of whose entries are equal to $\boldsymbol{G}$ (meaning that all the laboratories indeed are measuring the same quantity, "on average"), and with covariance matrix $\boldsymbol{S} = \boldsymbol{U} + \boldsymbol{V}$. Both $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n \times n$ symmetric matrices. The entries of the main diagonal of $\boldsymbol{U}$ are all equal to τ^2, and the off-diagonal entries are all zero, except those that correspond to the pairs of laboratories mentioned above: $0.351\tau^2$ or $0.234\tau^2$, respectively. $\boldsymbol{V}$ denotes a diagonal matrix with the $\{u_j^2\}$ in the main diagonal.

The question may reasonably be asked of whether the stated correlations are to be taken at face value, or whether there is some margin of doubt as to their values. If there is some uncertainty associated with them, then this can be recognized at least during the evaluation of $u(G)$ via the parametric statistical bootstrap by sampling from a suitably calibrated probability distribution of Fisher's z-transform of the correlation coefficient (Fisher, 1915, 1921).

The probability density to be maximized with respect to $\boldsymbol{G}$ and τ is

$$f(\boldsymbol{x}|G,\tau) = (2\pi)^{-n/2}|S^{-1}|^{1/2} \exp\left\{-1/2(\boldsymbol{x}-\boldsymbol{\mu})^\top S^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},$$

where $\top$ denotes matrix transposition, $\boldsymbol{x} = (x_1, \dots, x_n)^\top$ $\boldsymbol{\mu} = (G, \dots, G)^\top$ are column vectors, and $\boldsymbol{S}$ is as defined above, with inverse S^{-1}, and $|S^{-1}|$ denotes the determinant of its inverse.

The maximization was done numerically, under the constraints that both $\boldsymbol{G}$ and τ be non-negative, using function `nloptr` defined in the package of the same name for the R environment for statistical computing and graphics (Ypma, 2014; Johnson, 2015; R Core Team, 2015), using the "Subplex" algorithm (Rowan, 1990). According to the theory of maximum likelihood estimation (Wasserman, 2004), the results of the optimization can also be used to obtain an approximation for $u(G)$. The quality of this approximation generally tends to increase with increasing number n of laboratories.

Based on the data in Exhibit 35, and the modeling assumptions just described, the maximum likelihood estimate of $\boldsymbol{G}$ is $6.673\,81 \times 10^{-11}$ m^3 kg^{-1} s^{-2}, with approximate associated standard uncertainty $u(G) = 0.000\,31 \times 10^{-11}$ m^3 kg^{-1} s^{-2}. This consensus value and standard uncertainty are depicted in the same Exhibit 35. Obviously, the maximum likelihood estimate and the 2014 CODATA recommended value are statistically indistinguishable once the corresponding associated uncertainties are taken into account.

The standard deviation τ, of the laboratory effects $\lambda_1, \dots, \lambda_n$, also is of scientific interest because it quantifies the extent of the disagreement between the values measured by the different laboratories, above and beyond the differences that would be expected based only on the stated laboratory-specific standard uncertainties $\{u_j\}$.

The maximum likelihood estimate of τ is $0.001\,021\,5 \times 10^{-11}$ m^3 kg^{-1} s^{-2}, which is 3.8

times larger than the median of the $\{u_j\}$, suggesting that there may be very substantial sources of uncertainty still to be characterized that are responsible for that disagreement.

E22 Copper in Wholemeal Flour. The Analytical Methods Committee (1989) of the Royal Society of Chemistry lists the following determinations of the mass fraction of copper (expressed in µg/g) in wholemeal flour obtained under conditions of repeatability (VIM 2.20): 2.9, 3.1, 3.4, 3.4, 3.7, 3.7, 2.8, 2.5, 2.4, 2.4, 2.7, 2.2, 5.28, 3.37, 3.03, 3.03, 28.95, 3.77, 3.4, 2.2, 3.5, 3.6, 3.7, 3.7.

This Committee recommended that a Huber M-estimator of location (Huber and Ronchetti, 2009) be used instead of the simple arithmetic average when the determinations do not appear to be a sample from a Gaussian distribution, and indeed in this case the Anderson-Darling test rejects the hypothesis of Gaussian shape (Anderson and Darling, 1952).

Function huberM defined in R package robustbase (Rousseeuw et al., 2012), implements a robust alternative to the arithmetic average that yields both an estimate of that mass fraction and an evaluation of the associated standard uncertainty. (Note that among the arguments of the function huberM there is an adjustable parameter whose default value $k = 1.5$ may not be best in all cases.)

This function, applied with the default values of its arguments, produces 3.21 µg/g as an estimate of the measurand, and standard uncertainty 0.14 µg/g.

Since outliers may contain valuable information about the quantity of interest, it may be preferable to model them explicitly instead of down-weighing them automatically.

Function BESTmcmc defined in R package BEST (Kruschke, 2013; Kruschke and Meredith, 2013) implements a model-based Bayesian alternative: the data are modeled as a sample from a Student's t distribution with ν degrees of freedom, re-scaled to have standard deviation σ, and shifted to have mean equal to the measurand, using minimally informative prior distributions.

This Bayesian model is similar to a model used by Possolo (2012), and effectively selects the heaviness of the tails (quantified by ν), of the sampling distribution for the data, in a data-driven way. The corresponding posterior distribution for the mass fraction describes the associated uncertainty fully: the mean of this distribution is an estimate of the measurand, and its standard deviation is an evaluation of the associated standard uncertainty.

Function BESTmcmc, applied with the default values of its arguments, produces a posterior distribution for the mass fraction whose mean and standard deviation are 3.22 µg/g and 0.15 µg/g. (The posterior mean for ν was 1.8, suggesting very heavy tails indeed — Exhibit 36.)

Compare these with the conventional average, 4.28 µg/g, and standard error of the average $s/\sqrt{m} = 5.3/\sqrt{24} \approx 1.08$ µg/g, where s denotes the standard deviation of the $m = 24$ determinations. However, the results from the Bayesian analysis are in close agreement with the results of the classical robust analysis using the Huber procedure discussed above.

Coverage intervals can also be derived from the posterior distribution: for example, the interval ranging from 2.92 µg/g to 3.51 µg/g includes 95 % of the sample that BESTmcmc drew from the posterior distribution (via Markov Chain Monte Carlo sampling (Gelman et al., 2013)), hence is an approximate 95 % coverage for the mass fraction of copper.

E23 Tritium Half-Life. Lucas and Unterweger (2000, Table 2) list thirteen measurement

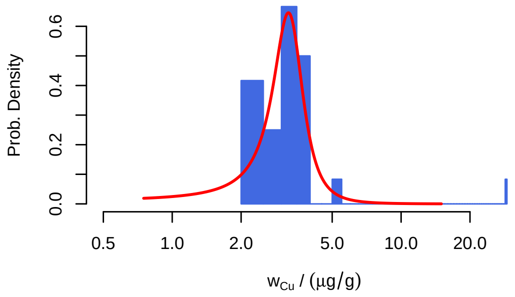


Exhibit 36: Histogram of the determinations of the mass fraction of copper in wholemeal flour, and Bayesian predictive density (red curve) that is a rescaled and shifted Student's *t* distribution with 1.8 degrees of freedom. (Note the logarithmic scale of the horizontal axis.)

results for the half-life $T_{½}$ of tritium, assembled as a result of a systematic review of the literature, reproduced in Exhibit 37.

A consensus estimate of $T_{½}$ may be produced based on the following observation equation that expresses the value of the half-life measured in study j as $T_j = T_{½} + \lambda_j + \varepsilon_j$, where λ_j denotes an effect specific to study j, and ε_j denotes measurement error, for $j = 1, \dots, 13$.

The study effects $\{\lambda_j\}$ and the measurement errors $\{\varepsilon_j\}$ are modeled as outcomes of independent Gaussian random variables, all with mean zero, the former with (unknown) standard deviation τ, and the latter with standard deviations equal to the corresponding standard uncertainties $\{u_j\}$ (Exhibit 37), which are assumed known.

It is the presence of the $\{\lambda_j\}$ that gives this model its name, *random effects model*, and that allows it to accommodate situations where the variability of the estimates $\{T_j\}$ exceeds what would be reasonable to expect in light of the associated uncertainties $\{u_j\}$. As we shall see below, such excess variability appears to be modest in this case.

This model may be fitted to the data in any one of several different ways. One of the most popular fitting procedures was suggested by DerSimonian and Laird (1986), and it is implemented in function `rma` defined in R package `metafor` (Viechtbauer, 2010; R Core Team,

		SOURCE	T_j /day	u_j
N	1947	Novick	4419	183
Je	1950	Jenks *et al.*	4551	54
J1	1951	Jones	4530	27
J5	1955	Jones	4479	11
P	1958	Popov	4596	66
M	1966	Merritt & Taylor	4496	16
J6	1967	Jones	4474	11
R	1977	Rudy	4501	9
S	1987	Simpson	4498	11
O	1987	Oliver	4521	11
A	1988	Akulov *et al.*	4485	12
B	1991	Budick *et al.*	4497	11
U	2000	Unterweger & Lucas	4504	9

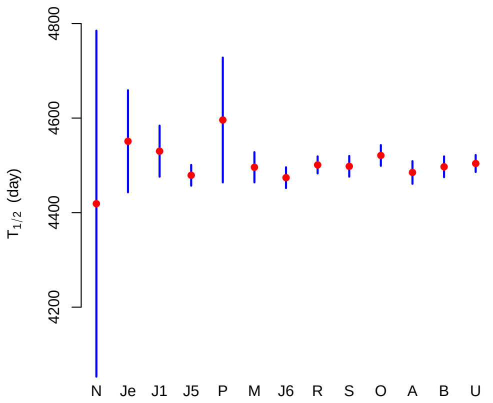


Exhibit 37: LEFT PANEL: Measurement results for the half-life of tritium, where T_j denotes the estimate of the half-life $T_{½}$, and u_j denotes the associated standard uncertainty, for $j = 1, \dots, 13$. RIGHT PANEL: The vertical (blue) line segments depict coverage intervals of the form $T_j \pm 2u_j$, and the (red) dots indicate the estimates of the half-life.

2015).

When rma is used including the optional adjustment suggested by Knapp and Hartung (2003), it produces 4497 day as consensus estimate of the half-life of tritium with associated standard uncertainty 5 day. This measurement result is statistically indistinguishable from the 4500 day, with associated standard uncertainty 8 day, recommended by Lucas and Unterweger (2000).

Function rma also produced the estimate $\widehat{\tau} = 10$ day of the standard deviation τ of the study effects $\{\lambda_j\}$. The heterogeneity metric I^2, suggested by Higgins and Thompson (2002), equals 35 %: this is the proportion of the total variability in the estimates of the study effects $\{\widehat{\lambda}_j\}$ that is attributable to differences between them (that is, heterogeneity), above and beyond study-specific measurement uncertainty. Therefore, there is modest heterogeneity in this case.

Lucas and Unterweger (2000) note that, in an evaluation such as they undertook, "the most difficult problem is to evaluate the uncertainty associated with each measurement in a consistent way". In other words, they are based on small, yet unspecified, numbers of degrees of freedom. To address this problem, they performed their own evaluation of the standard uncertainty associated with each measured value. If the original author's evaluation was neither larger than twice their evaluation, nor less than half as large, then they kept the original author's evaluation. Otherwise, they replaced the original author's u_j with theirs.

Lucas and Unterweger (2000) also state: "We can not emphasize strongly enough that estimated uncertainties have large uncertainties". Since the standard uncertainty $u(T_{½}) = 5$ day associated with the consensus value produced by the aforementioned DerSimonian-Laird procedure involves the unrealistic assumption that the $\{u_j\}$ are based on infinitely many degrees of freedom, it may well be that $u(T_{½}) = 5$ day is too optimistic.

In addition, the uncertainty associated with the estimate $\widehat{\tau}$ of the standard deviation of the study effects, which also figures in $u(T_{½})$, may not have been taken fully into account. Both issues may be addressed by performing a Monte Carlo evaluation of the uncertainty associated with the consensus value instead of relying on formulas that rest on assumptions that may be questionable.

To characterize the reliability of the study-specific uncertainties $\{u_j\}$, we need an assessment of the effective number of degrees of freedom associated with these standard uncertainties. Suppose that the $\{u_j\}$ are all based on the same number ν of degrees of freedom, that their common true value is σ, and that they differ from one another (and from σ) owing to the vagaries of sampling only. Together with the assumption that the data are like outcomes of Gaussian random variables, those suppositions imply that the $\{\nu u_j^2/\sigma^2\}$ should be like a sample from a chi-squared distribution with ν degrees of freedom, whose variance is 2ν.

Therefore, the variance of the $\{u_j^2/\sigma^2\}$ should be $2/\nu$. If we estimate σ by the median of the $\{u_j\}$ listed in in Exhibit 37, and then compute a robust estimate (the square of R's mad) of the variance of the ratios $\{u_j^2/\widehat{\sigma}^2\}$, we obtain 0.24, hence $\nu = 2/0.24 \approx 8$ is the effective number of degrees of freedom.

The uncertainty associated with $\widehat{\tau}$ may be characterized using the following representation of $\widehat{\tau}^2$ (Searle et al., 2006, 3.6-vii):

$$\widehat{\tau}^2 = \max\{0, \frac{((\nu+1)\tau^2 + \sigma^2)\chi^2_{n-1}}{(n-1)(\nu+1)} - \frac{\sigma^2\chi^2_{n\nu}}{n(\nu+1)\nu}\},$$

where $n = 13$ denotes the number of studies, and χ^2_{n-1} and $\chi^2_{n\nu}$ denote random variables with chi-squared distributions with $n-1$ and $n\nu$ degrees of freedom.

Alternatively, and possibly more accurately, the uncertainty associated with $\widehat{\tau}$ may be characterized by sampling from the probability distribution whose density is given in Equation (9) of Biggerstaff and Tweedie (1997), or using their Equation (6) together with the exact distribution of Cochran's heterogeneity statistic Q derived by Biggerstaff and Jackson (2008). These alternatives will not be pursued here.

To perform the Monte Carlo uncertainty evaluation repeat the following steps for $k = 1, \dots, K$, for a sufficiently large number of steps K:

1. Draw a sample value τ^2_k from the sampling distribution of $\widehat{\tau}^2$ specified above;
2. For each $j = 1, \dots, n$, compute $\sigma^2_{j,k} = \nu u^2_j / v^2_{j,k}$ where the $\{v^2_{j,k}\}$ denote independent chi-squared random variables with ν degrees of freedom;
3. For each $j = 1, \dots, n$, draw a sample value $T_{j,k}$ from a Gaussian distribution with mean $\widehat{T}_{½}$ (the estimate of the half-life produced by the DerSimonian-Laird procedure), and standard deviation $\sqrt{\tau^2_k + \sigma^2_{j,k}}$;
4. Apply the DerSimonian-Laird procedure to the $\{(T_{j,k}, \sigma_{j,k})\}$ to obtain an estimate $T^*_{½,k}$ of the half-life.

The standard deviation of $\{T^*_{½,1}, \dots, T^*_{½,K}\}$ is the Monte Carlo evaluation of the standard uncertainty associated with the DerSimonian-Laird estimate of the consensus value, taking into account the fact that the study-specific standard uncertainties are based on only finitely many degrees of freedom, and that the estimate $\widehat{\tau}$ of the standard deviation of the differences between studies is based on a fairly small number of degrees of freedom.

With $K = 10^6$, that standard deviation turned out to be $u(T_{½}) = 4.991$ day, while function `rma` produced $u(T_{½}) = 4.852$ day. Even though the Monte Carlo evaluation produces a slightly larger value, in this case the difference is inconsequential, both rounding to 5 day. It is good to know that this uncertainty evaluation is not particularly sensitive to the choice of the effective number of degrees of freedom, ν, associated with the $\{u_j\}$: had it been 1 instead of 8, then $u(T_{½})$ would have grown to only 6 day.

E24 Leukocytes. Measuring the number concentration of different types of white blood cells (WBC) is one of the most common procedures performed in clinical laboratories. The result is often based on the classification of 100 leukocytes into different types by microscopic examination. Fuentes-Arderiu and Dot-Bach (2009) report the counts listed in Exhibit 38, for a sample whose total number concentration of leukocytes was 3500 /µL.

To evaluate the uncertainty associated with the count in each class we should take into account the fact that an over-count in one class will induce an undercount in one or more of the other classes. Therefore, the counts should be modeled as outcomes of dependent random variables.

The multinomial probability distribution is one model capable of reproducing this behavior, and once it has been fitted to the data it may be used to produce coverage intervals for the proportions of leukocytes of the different types. The procedure proposed by Sison and Glaz

LEUKOCYTE	COUNT	c_{BE}	$U_{95\,\%}$	95 % CI
Neutrophils	63	0.066	14.0	(49, 77)
Lymphocytes	18	0.325	12.5	(5, 29)
Monocytes	8	0.55	8.5	(0, 17)
Eosinophils	4	0.688	5.0	(0, 10)
Basophils	1	2.632	3.0	(0, 6)
Myelocytes	1	1.325	2.0	(0, 4)
Metamyelocytes	5	0.696	6.5	(0, 13)

Exhibit 38: Number in each of seven classes after classification of 100 leukocytes in a blood smear (COUNT), coefficient of variation (c_{BE}) reflecting between-examiner reproducibility for each class as determined by Fuentes-Arderiu et al. (2007), expanded uncertainty computed using the Monte Carlo method ($U_{95\,\%}$), and 95 % coverage intervals for true counts.

(1995), implemented in R function `multinomialCI` defined in package `MultinomialCI` (Villacorta, 2012), produces coverage intervals for those proportions with any specified coverage probability.

If, in particular, the function is used to produce 68 % coverage intervals, then one half of the lengths of the resulting intervals may be interpreted as evaluations of the standard uncertainties u_{M} associated with the proportions corresponding to the multinomial model. For the proportion of neutrophils this interval ranges from 0.600 to 0.663, hence the standard uncertainty associated with the number of neutrophils in a sample of 100 leukocytes is is $100(0.6627 - 0.6000)/2 = 3.13$.

However, the other identified source of uncertainty, between-examiner reproducibility, also must be taken into account. In a separate study, Fuentes-Arderiu et al. (2007) determined the coefficients of variation (ratios of standard deviations to averages) for the different classes, that are attributable to lack of reproducibility, also listed in Exhibit 38. For example, the standard uncertainty u_{BE} corresponding to this source for the number of neutrophils is $63 \times 0.066 = 4.16$.

The conventional way of combining the contributions from these two sources of uncertainty, whose standard uncertainties are u_{M} and u_{BE}, is in root sum of squares, which for the number of neutrophils would yield $\sqrt{3.13^2 + 4.16^2} = 5.21$. This manner of combining these contributions presupposes that deviations to either side of the true count are equally likely, for each of these two sources of uncertainty. While this may be reasonable for counts that are far away from 0 by comparison with the corresponding values of u_{M} and u_{BE}, it is quite unreasonable for counts like the 1 for basophils, for which u_{M} equals 2.13 and u_{BE} equals 2.632.

An alternative, more realistic evaluation will take into account the constraint captured in the multinomial model: that the counts must add to 100, and that all counts must be greater than or equal to 0 irrespective of how large the associated uncertainties may be.

The following Monte Carlo procedure is one way of implementing this alternative evaluation, and involves repeating the following steps a sufficiently large number K of times, where $p = (63, 18, 8, 4, 1, 1, 5)/100$ is the vector of proportions of the different classes of leukocytes, and $n = 7$ denotes the number of classes, for $k = 1, \dots, K$:

1. Draw a vector $x_k = (x_{1,k}, \dots, x_{n,k})$ with n counts from the multinomial distribution determined by probabilities p and size 100.

2. Draw a sample value $b_{j,k}$ from a Gaussian distribution with mean 0 and standard de-

viation $u_{\mathrm{BE},j}$, representing the measurement error corresponding to between-examiner variability, for class $j = 1, \dots, n$.

3. Compute $s_{j,k} = \max(0, x_{j,k} + b_{j,k})$, which forces the Monte Carlo sample count for class j to be non-negative, for $j = 1, \dots, n$;
4. Define $y^*_{j,k}$ as the value of $s_{j,k}/(s_{1,k} + \cdots + s_{n,k})$ after rounding to the nearest integer.

Next, and for each class $j = 1, \dots, n$, compute one half of the difference between the 97.5th and 2.5th percentiles of the Monte Carlo sample of values $\{y^*_{j,1}, \dots, y^*_{j,K}\}$ that have been drawn from the distribution of the count for this class, to obtain approximate expanded uncertainties $U_{95\,\%}$. The 2.5th and 97.5th percentiles (possibly rounded to the nearest integer) are the end-points of 95 % coverage intervals for the true counts in the different classes (Exhibit 38).

E25 Yeast Cells. William Sealy Gosset (*Student*) used a hemacytometer to count the number of yeast cells in each of 400 square regions on a plate, arranged in a 20×20 grid whose total area was 1 mm^2, and reported the results as the numbers of these regions that contained 0, 1, 2, …, yeast cells, as follows: $(0,0)$, $(1,20)$, $(2,43)$, $(3,53)$, $(4,86)$, $(5,70)$, $(6,54)$, $(7,37)$, $(8,18)$, $(9,10)$, $(10,5)$, $(11,2)$, $(12,2)$. For example, there were no regions with no cells, twenty regions contained exactly one cell each, and forty-three regions contained exactly two cells each. The purpose is to estimate the mean number of yeast cells per 0.0025 mm^2 region, in preparations made similarly to this plate, as described by Student (1907).

The measurement model describes these counts $z_1, \dots, z_m$ as realized values of $m = 400$ independent random variables with a common Poisson distribution with mean λ, which is the measurand. This model is commonly used to describe the variability of the number of occurrences of an event that results from the cumulative effect of many improbable causes (Feller, 1968, XI.6b), and it models acceptably well the dispersion of these data. In fact, the sample mean (4.68) and the sample variance (4.46) of the observed counts are numerically close as expected under the Poisson model, and a conventional chi-squared goodness-of-fit test also supports the assumption of Poissonness.

A Bayesian estimate of λ may be derived from a posterior distribution (Possolo and Toman, 2011) computed using the likelihood function corresponding to the Poisson model, and the probability density suggested by Jeffreys (1961) that describes the absence of information about the value of λ prior to the experiment. This density is proportional to $1/\sqrt{\lambda}$. Since its integral from zero to infinity diverges, it is an improper prior probability density. However, the corresponding posterior distribution is proper and its density, q, can be calculated explicitly by application of Bayes's rule, where $s = z_1 + \cdots + z_m = m\overline{z}$:

$$q(\lambda | z_1, \dots, z_m) = \frac{\dfrac{\lambda^s \exp(-\lambda m)}{z_1! \dots z_m!} \dfrac{1}{\sqrt{\lambda}}}{\displaystyle\int_0^{+\infty} \dfrac{l^s \exp(-lm)}{z_1! \dots z_m!} \dfrac{1}{\sqrt{l}} \mathrm{d}l} = \frac{m^{s+½}}{\Gamma(s + ½)} \lambda^{s-½} \exp(-\lambda m).$$

This is the probability density of a gamma distribution with shape $m\overline{z} + ½$ and scale $1/m$, hence with mean $\widehat{\lambda} = \overline{z} + 1/(2m) = 4.68$ and standard deviation $u(\lambda) = \sqrt{\overline{z}/m + 1/(2m^2)}$

$= 0.11$. Exhibit 39 depicts the corresponding probability density, and a 95 % coverage interval for λ, ranging from 4.47 to 4.90.

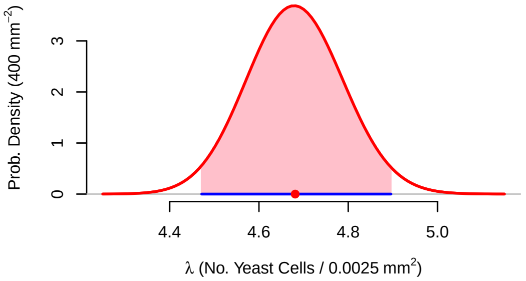


Exhibit 39: Probability density of the posterior distribution of λ given the data, and a 95 % coverage interval for λ (mean number of yeast cells per 0.0025 mm^2 region).

E26 Refractive Index. A solid glass prism has been placed on the table of a refractometer with the edge where its refracting faces meet parallel to the rotation axis of the table, and a beam of monochromatic light has been made to traverse it on a plane perpendicular to that axis (Exhibit 40).

As the prism is rotated, the angle between the direction of the beam when it enters the prism and its direction when it exits the prism varies. When this deviation angle reaches its minimum value at δ, the following relationship (measurement equation) holds between the true values of the prism's apex angle α, and of the refractive indexes n_{G} and n_{A} of the glass and of the air the prism is immersed in (Jenkins and White, 1976, §2.5):

$$n_{\mathrm{G}} = n_{\mathrm{A}} \frac{\sin\left(\frac{\alpha+\delta}{2}\right)}{\sin\left(\frac{\alpha}{2}\right)}.$$

The only sources of uncertainty recognized and propagated in this example are: (i) lack of repeatability of replicated determinations of the apex angle α and of the minimum deviation angle δ; and (ii) measurement uncertainty of the refractive index of the air, n_{A}. The contributions from the two sources in (i) were evaluated using Type A methods, and the contribution from (ii) was evaluated using a Type B method.

The refractive index of air was estimated using a modified Edlén's formula (Edlén, 1966; Stone and Zimmerman, 2011) as $n_{\mathrm{A}} = 1.000\,264\,3$, with standard measurement uncertainty $u(n_{\mathrm{A}}) = 0.000\,000\,5$ (Fraser and Watters, 2008, Table 2).

The six replicates of the minimum deviation angle $d_1, \dots, d_6$ were 38.661 169°, 38.661 051°, 38.660 990°, 38.660 779°, 38.661 075°, and 38.661 153°. The sixteen replicates $a_1, \dots, a_{16}$ of the prism's apex angle were:

60.007 314°	60.007 169°	60.007 367°	60.006 969°	60.006 972°	60.006 586°	60.007 172°	60.007 017°
60.006 533°	60.006 242°	60.006 358°	60.006 308°	60.006 369°	60.006 297°	60.005 806°	60.006 333°

Measurement Equation. A conventional method to estimate the measurand consists of using the measurement equation $n_{\mathrm{G}} = n_{\mathrm{A}} \sin\big((\alpha + \delta)/2\big) / \sin(\alpha/2)$ with $\alpha = \bar{a} = (a_1 + \dots + a_{16})/16$, $\delta = \bar{d} = (d_1 + \dots + d_6)/6$, and $n_{\mathrm{A}} = 1.000\,264\,3$. The choice of averages is validated by the fact that both the $\{a_i\}$ and the $\{d_j\}$ may be regarded as samples from Gaussian distributions, based on the Shapiro-Wilk goodness-of-fit test (Shapiro and Wilk, 1965). Thus $n_{\mathrm{G}} = 1.517\,287$, and the conventional formula for uncertainty propagation yields $u(n_{\mathrm{G}}) = 0.000\,002$.

Measurement uncertainty can also be evaluated by application of the Monte Carlo method of the GUM-S1, based on these assumptions:

(i) $\sqrt{16}(\bar{a} - \alpha)/s_a$ is like an outcome of a Student's t random variable with 15 degrees of freedom, where $s_a = 0.000\,008\,1°$ is the standard deviation of the sixteen $\{a_i\}$;

(ii) $\sqrt{6}(\bar{d} - \delta)/s_d$ is like an outcome of a Student's t random variable with 5 degrees of freedom, where $s_d = 0.000\,002\,5°$ is the standard deviation of the six $\{d_j\}$.

(iii) The $\{a_i\}$ and the $\{d_j\}$ are independent.

This evaluation reproduces the single significant digit given above for $u(n_{\mathrm{G}})$. In addition, it also provides a sample drawn from the distribution of the measurand whose 2.5th and 97.5th percentiles are the endpoints of the following 95 % coverage interval for the true value of the refractive index: $(1.517\,284, 1.517\,290)$. Exhibit 40 shows an estimate of the probability density of the distribution of the measurand.

Observation Equations. Yet another method to estimate the measurand and to evaluate the associated uncertainty is based on the following observation equations (which are to be understood modulo 360° because they involve angles): $a_i = \alpha + r_i$, for $i = 1, \dots, 16$, and $d_j = H(\nu_{\mathrm{G}}, \nu_{\mathrm{A}}, \alpha) + s_j$, for $j = 1, \dots, 6$. The $\{r_i\}$ and the $\{s_j\}$ denote non-observable measurement errors, and the function H is defined by $H(\nu_{\mathrm{G}}, \nu_{\mathrm{A}}, \alpha) = 2 \arcsin(\nu_{\mathrm{G}} \sin(\alpha/2)/\nu_{\mathrm{A}}) - \alpha$. The specification of this statistical model is completed by assuming that the $\{r_i\}$ and the $\{s_j\}$ are like outcomes of independent, Gaussian random variables, all with mean zero, the former with standard deviation σ_α, the latter with standard deviation σ_δ.

This model may be fitted by the method of maximum likelihood, which in this case involves solving a constrained non-linear optimization problem. Employing the Nelder-Mead method (Nelder and Mead, 1965) yields the estimate $\widehat{n}_{\mathrm{G}} = 1.517\,287$, which is identical to the estimate derived from the approach based on the measurement equation. The Monte Carlo evaluation consistent with these observation equations is the so-called parametric statistical bootstrap (Efron and Tibshirani, 1993), and its results reproduce the values indicated above both for the standard uncertainty and for the 95 % coverage interval.

E27 Ballistic Limit of Body Armor. The *ballistic limit* υ_{50} of a particular type of bullet-proof vest for a particular type of bullet is the velocity at which the bullet penetrates the vest with 50 % probability. To measure υ_{50}, several bullets of different velocities are fired at identical vests under standardized conditions, for example as specified by OLES (2008), and for each of them the result is recorded as a binary (nominal) outcome indicating whether the vest stopped the bullet or not. The input variables are bullet velocity and this binary outcome.

These are the results of a particular test conducted at NIST (Mauchant et al., 2011) that involved firing $m = 15$ bullets at several identical vests: $(374.5, 0)$, $(415, 1)$, $(407, 1)$, $(387.5, 1)$, $(372.5, 0)$, $(399.5, 1)$, $(391, 0)$, $(408.5, 0)$, $(427, 0)$, $(446, 1)$, $(441, 1)$, $(422, 0)$, $(432, 0)$, $(451, 1)$, $(443, 1)$. The first value in each pair is the bullet velocity (expressed in m/s), and the second indicates whether the bullet did (1) or did not (0) penetrate the vest.

A possible measurement model for υ_{50} involves an observation equation and a measurement equation. The observation equation in turn comprises two parts. This first part is a Bernoulli

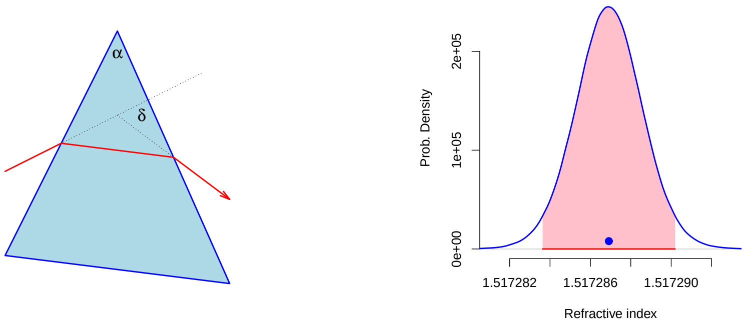


Exhibit 40: LEFT PANEL: Cross-section of a triangular prism of apex angle α, standing on a plane parallel to the plane of the figure, and the path of a monochromatic light beam that enters the prism from the left and exits it on the right, in the process undergoing a total deviation δ. RIGHT PANEL: Estimate of the probability density that characterizes the measurement uncertainty of the refractive index of the glass. The (blue) dot marks the average of the Monte Carlo sample of size $K = 10^6$. Since the lightly shaded (pink) region comprises 95 % of the area under the curve, the thick, horizontal (red) line indicates a 95 % coverage interval for the measurand.

model for bullet penetration, which states that the results from different shots are like the outcomes of independent tosses of different coins, the coin corresponding to a bullet of velocity v having probability (v) of "heads", denoting penetration. The second part is a logistic regression model for these probabilities, $\log((v)/(1 - (v))) = \alpha + \beta v$, where α and β are parameters to be estimated. The measurement equation is $v_{50} = -\alpha/\beta$.

Fitting the model to the data above by the method of maximum likelihood produces $\widehat{\alpha} = -14.5$ and $\widehat{\beta} = 0.035\,34$ s/m, hence $\widehat{v}_{50} = -\widehat{\alpha}/\widehat{\beta} = 410$ m/s. Exhibit 41 depicts the data and the fitted logistic regression function. The maximum likelihood procedure also provides evaluations of the standard uncertainties and covariance for $\widehat{\alpha}$ and $\widehat{\beta}$. Application of the NUM then yields $u(v_{50}) = 16$ m/s.

A parametric statistical bootstrap (Efron and Tibshirani, 1993) could be used instead for the uncertainty evaluation. The idea is to simulate values of binary random variables $B_1, \dots, B_m$ to synthesize data $(v_1, B_1), \dots, (v_m, B_m)$, where $v_1 = 374.5$ m/s, …, $v_m = 443$ m/s are kept fixed at the actual bullet velocities achieved in the experiment, and to use these simulated data to produce an estimate of the ballistic limit. B_i has a Bernoulli probability distribution with probability of "success" (that is, penetration) $\widehat{}(v_i)$ such that $\log(\widehat{}(v_i)/[1 - \widehat{}(v_i)]) = \widehat{\alpha} + \widehat{\beta} v_i$, for $i = 1, \dots, m$.

Repeating this process a large number $K = 10\,000$ of times produces a sample of estimates that may be used to characterize the associated uncertainty. Since $v_{50} = -\alpha/\beta$ is a ratio, and some of the Monte Carlo sample values for the denominator may be very close to 0, the corresponding sample drawn from the probability distribution of v_{50} may include values that lie very far from its center. For this reason, we use the scaled median absolute deviation from the median (`mad`) to obtain an estimate of the standard deviation of that distribution that is robust to such extreme values: $u(v_{50}) = 17$ m/s. A 95 % coverage interval for v_{50} ranged from 360 m/s to 460 m/s.

E28 Atomic Ionization Energy. NIST Standard Reference Database 141 (Kotochigova et al., 2011) includes results of *ab initio* local-density-functional calculations of total ener-

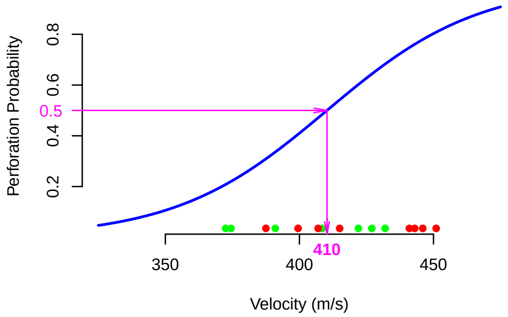


Exhibit 41: Logistic regression model fitted to the results of a test to measure the ballistic limit of a bullet-proof vest. The red dots indicate the bullet velocities that achieved penetration, and the green dots those that did not.

gies for the ground-state configurations of all atoms and singly-charged cations with atomic number $Z \leqslant 92$ (Kotochigova et al., 1997a,b).

Four standard approximations were used: (i) the local-density approximation (LDA); (ii) the local-spin-density approximation (LSD); (iii) the relativistic local-density approximation (RLDA); and (iv) the scalar-relativistic local-density approximation (ScRLDA). For example, these approximations yield the following estimates of the first ionization energy of ^{20}Ca: 6.431 274 52 eV (LDA), 6.210 943 94 eV (LSD), 6.453 451 8 eV (RLDA), and 6.453 479 01 eV (ScRLDA), where $1\,\text{eV} = 1.602\,176\,57 \times 10^{-19}$ J. The corresponding value that was measured experimentally is 6.113 155 20 eV with standard uncertainty 0.000 000 25 eV (Miyabe et al., 2006; Kramida et al., 2013).

Exhibit 42 lists values of the relative error of the LDA, LSD, and RLDA approximations used in the local-density-functional calculations of the first ionization energy of the alkaline earth metals: beryllium, magnesium, calcium, strontium, barium, and radium. (For these elements, the results of ScRLDA are almost indistinguishable from the results of RLDA, hence they are not shown separately.) Each of these relative errors is of the form $\varepsilon = (E_c - E)/E$, where E_c denotes the estimate obtained via a first-principles calculation (from Kotochigova et al. (2011)) and E denotes the corresponding value measured experimentally (Kramida et al., 2013).

The measurand is the standard deviation τ of the portion of the variability of the relative errors $\{\varepsilon_{ij}\}$ that is attributable to differences between LDA, LSD, and RLDA. The corresponding measurement model is the observation equation $\varepsilon_{ij} = \alpha_i + \beta_j + \delta_{ij}$, where ε_{ij} denotes the relative error for element i and approximation j, α_i denotes the effect of element i, β_j denotes the effect of approximation j, and δ_{ij} is a residual.

This model is a *mixed effects* model (Pinheiro and Bates, 2000), which is a generalization of the laboratory effects model discussed by Toman and Possolo (2009, 2010). Here the $\{\alpha_i\}$ represent "fixed" effects and the $\{\beta_i\}$ represent "random" effects. The former express differences between the elements with regards to the accuracy of the *ab initio* calculations. The latter are modeled as a sample from a Gaussian distribution with mean 0 and standard deviation τ. The "errors" $\{\delta_{ij}\}$ are regarded as a sample from a Gaussian distribution with mean 0 and standard deviation σ, which quantifies the intrinsic inaccuracy of the approximation methods.

The model was fitted using function `lme` defined in R package `nlme` (Pinheiro et al., 2014). None of the estimates of the element effects $\{\alpha_i\}$ differ significantly from 0. The estimate of the standard deviation that reflects differences between computational approximations is $\widehat{\tau} = 0.03$ eV, and the estimate of the standard deviation that characterizes the within-

Z		LDA	LSD	RLDA
4	Be	0.02019569	−0.03183248	0.02024823
12	Mg	0.05340336	0.00976185	0.05467029
20	Ca	0.05203850	0.01599646	0.05566630
38	Sr	0.04773201	0.01470489	0.06270221
56	Ba	0.02748393	−0.00244948	0.06102522
88	Ra	−0.03315584	−0.06014852	0.06543755

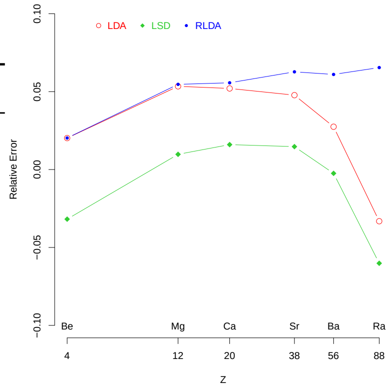


Exhibit 42: Values and graphical representation of the relative error of three approximations used in local-density-functional calculations of the first ionization energy of the alkaline earth metals. Each of these values is of the form $\varepsilon = (E_c - E)/E$, where E_c denotes an estimate obtained via an *ab initio* calculation (Kotochigova et al., 2011) and E denotes the corresponding value measured experimentally (Kramida et al., 2013).

element residuals is $\widehat{\sigma} = 0.021$ eV. Since these estimates are comparable, and in fact are not significantly different once their associated uncertainties are taken into account (evaluated approximately using function `intervals` defined in R package `nlme`), the conclusion is that, for the alkaline earth metals at least, the dispersion of values attributable to differences between computational approximations is comparable to the intrinsic (in)accuracy of the individual approximation methods.

E29 Forensic Glass Fragments. Evett and Spiehler (1987) point out that it is possible to determine the refractive index and chemical composition of even a very small glass fragment, as may be found in the clothing of a suspect of smashing a window to gain illicit access to a home or car. A forensic investigation may then compare the refractive index and chemical composition of the fragment of unknown provenance against a reference collection of samples of known type for which the same properties have been measured.

Evett and Spiehler (1987) describe a reference collection that was assembled by the Home Office Forensic Science Laboratory in Birmingham, United Kingdom, comprising 214 glass samples of known type with measured values of the refractive index and of the mass fractions of oxides of the major elements (Na, Mg, Al, Si, K, Ca, Ba, Fe). This is the *Glass Identification Data Set* in the Machine Learning Repository of the University of California at Irvine (Bache and Lichman, 2013), also available in object `glass` of the R package `mda` (Hastie et al., 2013).

The glass samples in this collection belong to the following types (with the number of corresponding samples between parentheses): float processed building windows (70), non-float processed building windows (76), float processed vehicle windows (17), containers (13), tableware (9), and headlamps (29). Modern windows (of buildings and vehicles) are made of glass that, while molten, was poured onto a bed of molten metal. This process, developed in the 1950s, yields glass sheets of very uniform thickness and superior flatness (Pilkington, 1969).

Consider a classifier that, given values of the refractive index, and of the mass fractions of the oxides of sodium, magnesium, aluminum, silicon, potassium, calcium, barium, and iron (quantitative inputs), produces an estimate of glass type (qualitative output) as one of the six types described above. The classifier we will consider was built using mixture discriminant analysis (Hastie et al., 2009) as implemented in function `mda` of the R package of the same name (Hastie et al., 2013). Exhibit 43 depicts the data in the space of "canonical variables" computed by `mda`.

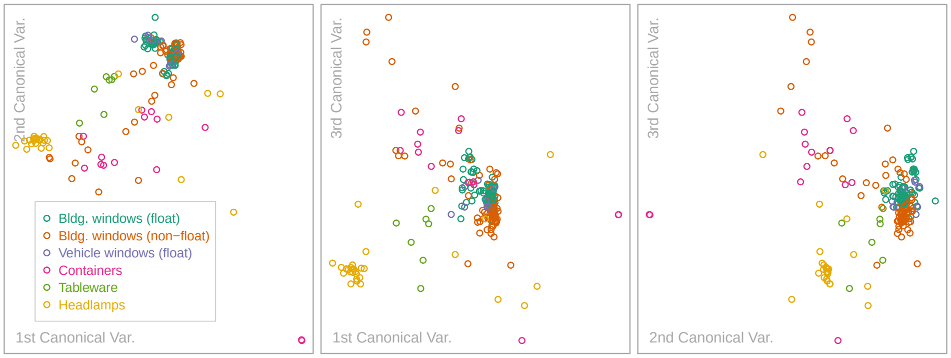


Exhibit 43: Projection of the glass data onto the three "canonical" variables that account for 87 % of the variability in the data.

Given values of the inputs for a particular glass fragment, the classifier computes a probability distribution over the set of possible types, and assigns the fragment to the type that has the highest probability. Suppose that a glass fragment of unknown provenance has refractive index 1.516 13 and mass fractions of the oxides of the major elements (%): 13.92, 3.52, 1.25, 72.88, 0.37, 7.94, 0, and 0.14. The classifier produces the following probability distribution for the provenance of the fragment: building windows (float glass), 0.36; building windows (non-float glass), 0.56; vehicle windows (float glass), 0.08; containers, 0.00; tableware, 0.00; headlamps, 0.00.

Therefore, with 36 % + 56 % = 92 % probability the fragment is from a building window, and it is more likely to be from an old building (non-float glass) than from a modern building (float glass). The corresponding value of the output is "building windows (non-float glass)" because it has the highest probability, but this assignment is clouded by the considerable uncertainty conveyed by that probability distribution.

Similarly to how the entropy was considered in Exhibit 8 on Page 37, this uncertainty may be quantified using the entropy of the probability distribution over the six types of glass that was produced by the classifier: $0.89 = -(0.36\log(0.36) + 0.56\log(0.56) + 0.08\log(0.08))$. Since the entropy of a Gaussian distribution with standard deviation σ is $\frac{1}{2}\log(2\quad e) + \log\sigma$, one may argue that $\exp(0.89 - \frac{1}{2}\log(2\quad e)) = 0.59$ is an analog of the standard uncertainty. However, if the output of the classifier were to be used as input to other measurements, then the associated uncertainty should be propagated using the full distribution in the context of the Monte Carlo method, as was done for the Damerau–Levenshtein distance in Example E6.

The performance of the classifier may be evaluated using *leave-one-out cross-validation* (Mosteller and Tukey, 1977), as follows: for each glass sample in turn, build a classifier using the data for the other samples only, and use it to predict the type of the glass sample left out. The overall error rate is then estimated by the proportion of glass samples that were misclassified: this proportion was 30 % in this case.

E30 Mass of W Boson. The W boson is one of the elementary particles that mediates the weak interaction, and plays a role in some nuclear reactions, for example in the beta decay of tritium to helium, which is used in applications of radio-luminescence, for example in "tritium tubes" used to mark the hours on the faces of some watches.

Exhibit 44 lists and depicts the measurement results quoted by Olive and Particle Data Group (2014, Page 561), which have been obtained by the LEP Electroweak Working Group (involving the ALEPH, DELPHI, L3, and OPAL collaborations) (The ALEPH Collaboration et al., 2013) and by the Tevatron experiments (CDF and D0 collaborations) (CDF Collaboration and D0 Collaboration, 2013). The same Exhibit also indicates the estimate of the mass m_W of the W boson, and associated standard uncertainty, as computed by Olive and Particle Data Group (2014), and their counterparts (labeled "DL") based on a laboratory random effects model.

The laboratory random effects model is an observation equation that represents the value of the mass of the W boson measured by laboratory j as $m_{W,j} = m_W + \lambda_j + \varepsilon_j$, for $j = 1, \dots, n$, where $n = 6$ is the number of measurement results, m_W denotes the true value of the mass of the W boson, λ_j denotes an effect specific to experiment j, and ε_j denotes measurement error.

The laboratory effects $\{\lambda_j\}$ and the measurement errors $\{\varepsilon_j\}$ are modeled as outcomes of independent Gaussian random variables, all with mean zero, the former with (unknown)

	m_W	$u(m_W)$
COLLABORATION	GeV/c^2	
ALEPH	80.440	0.051
DELPHI	80.336	0.067
L3	80.270	0.055
OPAL	80.415	0.052
CDF	80.387	0.019
D0	80.375	0.023

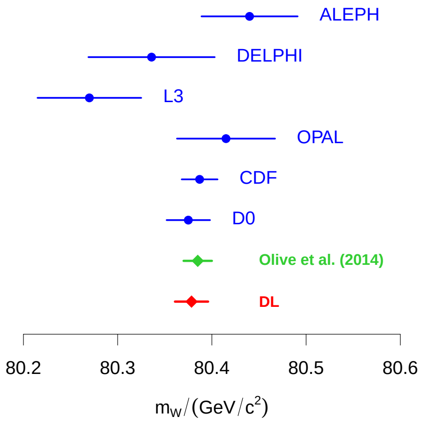


Exhibit 44: Measurement results for the mass of the W boson obtained by the LEP Electroweak Working Group (ALEPH, DELPHI, L3, OPAL) (The ALEPH Collaboration et al., 2013) and by the Tevatron experiments (CDF and D0) (CDF Collaboration and D0 Collaboration, 2013), summarized in Olive and Particle Data Group (2014, Page 561), where c denotes the speed of light in vacuum and $1\ \mathrm{GeV}/c^2 = 1.782\,662 \times 10^{-27}$ kg (left panel), and the estimates and uncertainty evaluations produced by Olive and Particle Data Group (2014), and by application of the DerSimonian-Laird procedure of meta-analysis (DerSimonian and Laird, 1986) (right panel). The measurement results are depicted in blue, with a dot indicating the measured value, and the horizontal line segment representing the interval $m_{W,j} \pm u(m_{W,j})$, for $j = 1, \dots, 6$.

standard deviation τ, the latter with standard deviations equal to the corresponding standard uncertainties $\{u_j(m_\mathrm{W})\}$ that are listed in Exhibit 44.

It is the presence of the $\{\lambda_j\}$ that gives this model its name, *random effects model*, and that allows it to accommodate situations where the variability of the estimates $\{m_{\mathrm{W},j}\}$ exceeds what would be reasonable to expect in light of the associated uncertainties $\{u_j(m_\mathrm{W})\}$. In this case, such excess variability appears to be modest because the standard deviation of the between-laboratory variability is estimated as $\widehat{\tau} = 0.019\,\mathrm{GeV}/c^2$, which is quite comparable in size to the $\{u_j(m_\mathrm{W})\}$ listed in Exhibit 44.

The random effects model may be fitted to the data in any one of several different ways. One of the most popular fitting procedures was suggested by DerSimonian and Laird (1986), and it is implemented in function `rma` defined in R package `metafor` (Viechtbauer, 2010; R Core Team, 2015).

The resulting estimate, $\widehat{m}_W = 80.378\,\mathrm{GeV}/c^2$ (where c denotes the speed of light in vacuum), and the associated standard uncertainty $0.018\,\mathrm{GeV}/c^2$, are depicted in Exhibit 44. Since $\widehat{m}_W = 80.378\,\mathrm{GeV}/c^2 = 1.4329 \times 10^{-25}\,\mathrm{kg}$, we conclude that the W boson is about 86 times more massive than a proton. The corresponding values computed by Olive and Particle Data Group (2014) are $80.385\,\mathrm{GeV}/c^2$ and $0.015\,\mathrm{GeV}/c^2$. Taking into account these uncertainties, it is obvious that the two consensus values are not significantly different statistically.

The DerSimonian-Laird procedure regards the $\{u_j(m_\mathrm{W})\}$ as if they were based on infinitely many degrees of freedom, and also fails to take into account the small number of degrees of freedom ($n - 1 = 5$ in this case) that the estimate of the inter-laboratory standard deviation τ is based on. This second shortcoming is mitigated by applying an adjustment suggested by Knapp and Hartung (2003), and the results given above reflect this.

A Monte Carlo evaluation of the uncertainty associated with the DerSimonian-Laird estimate may be performed by taking the following steps.

1. Model the estimate of τ^2 produced by R function `rma` as an outcome of a random variable with a lognormal distribution with mean equal to the estimated value $\widehat{\tau}^2 = 0.019\,\mathrm{GeV}/c^2$, and with standard deviation set equal to the estimate of the standard error of $\widehat{\tau}^2$ produced by `rma`, computed as explained by Viechtbauer (2007), $u(\tau^2) = 0.00104\,\mathrm{GeV}/c^2$.

2. Compute an effective number of degrees of freedom ν to associate with the $\{u(m_{\mathrm{W},j})\}$ to recognize, albeit coarsely, that they are based on finitely many numbers of degrees of freedom. We do this motivated by the following fact: if s is the standard deviation of a sample of size $\nu + 1$ drawn from a Gaussian distribution with standard deviation σ, then the variance of $\nu s^2/\sigma^2$ equals 2ν. Supposing that $u_1(m_\mathrm{W}), \dots, u_n(m_\mathrm{W})$ are like standard deviations of Gaussian samples all of the same size and with the same standard deviation σ, it follows that υ^2, the sample variance of $u_1^2(m_\mathrm{W}), \dots, u_n^2(m_\mathrm{W})$, should be $2\sigma^4/\nu$ approximately. Replacing σ by $\widetilde{\sigma} = \mathrm{median}\{u_j(m_\mathrm{W})\}$ leads to $\nu = 2\widetilde{\sigma}^4/\upsilon^2 = 5.64$.

3. Select a sufficiently large integer K and then repeat the following steps for $k = 1, \dots, K$:

   (a) Draw a value τ_k^2 from the lognormal probability distribution associated with $\widehat{\tau}$;

   (b) Draw a value $w_{j,k}^2$ from a chi-squared distribution with ν degrees of freedom, and compute $\sigma_{j,k} = \left(\nu u^2(m_{\mathrm{W},j})/w_{j,k}^2\right)^{½}$, for $j = 1, \dots, n$;

(c) Draw a value $\lambda_{j,k}$ from a Gaussian distribution with mean 0 and standard deviation τ_k for $j = 1, \dots, n$;

(d) Draw a value $\varepsilon_{j,k}$ from a Gaussian distribution with mean 0 and standard deviation $\sigma_{j,k}$, for $j = 1, \dots, n$;

(e) Compute $m_{\mathrm{W},j,k} = \widehat{m}_W + \lambda_{j,k} + \varepsilon_{j,k}$, for $j = 1, \dots, n$;

(f) Compute the DerSimonian-Laird estimate $m^*_{\mathrm{W},k}$ of m_W based on the n Monte Carlo measurement results $(m_{\mathrm{W},1,k}, \sigma_{1,k}), \dots, (m_{\mathrm{W},n,k}, \sigma_{n,k})$.

A Monte Carlo sample of size $K = 1 \times 10^5$ drawn from the distribution of the mass of the W boson as just described had standard deviation with the same two significant digits that R function rma produced for $u(m_\mathrm{W})$, thus lending credence to that uncertainty evaluation. A 95 % coverage interval derived from the Monte Carlo sample ranges from 80.343 GeV/c^2 to 80.414 GeV/c^2, while its counterpart produced by rma ranges from 80.333 GeV/c^2 to 80.424 GeV/c^2.

E31 Milk Fat. Exhibit 45 lists values of fat concentration in samples of human milk determined by two measurement methods, and shows a Bland-Altman plot of these data: one method is based on the measurement of glycerol released by enzymatic hydrolysis of triglycerides (Lucas et al., 1987), the other is the Gerber method (Badertscher et al., 2007).

Bland and Altman (1986) point out that a very high correlation between the paired measured values is a misleading indication of agreement between two measurement methods because a perfect correlation only indicates that the value measured by one method is a linear function of the value measured by the other, not that the corresponding measured values are identical. The correlation coefficient for these two sets of measured values is 0.998.

A paired t-test indicates that the mean difference does not differ significantly from zero. However, this, too, falls short of establishing equivalence (or, interchangeability) between the two measurement methods. If the paired samples are of small size, then there is a fair chance that a statistical test will fail to detect a difference that is important in practice. And if they are large, then a statistical test very likely will deem significant a difference that is irrelevant in practice (Carstensen, 2010).

For these reasons, Bland and Altman (1986) suggest that graphical methods may be particularly informative about the question of agreement between methods. This being the most often cited paper in the *Lancet* indicates the exceptional interest that measurement issues enjoy in medicine.

The Bland-Altman plot in Exhibit 45 shows how the difference between the paired measured values varies with their averages (Altman and Bland, 1983; Bland and Altman, 1986). Except for the inclusion of *limits of agreement* (the average of the differences between paired measured values plus or minus twice the standard deviation of the same differences), the Bland-Altman plot is the same as Tukey's mean-difference plot.

In this case, the difference between the methods tends to be positive for small values of the measurand, and negative for large values. Exhibit 46 shows a Bland-Altman plot that recognizes this trend. Function BA.plot from R package MethComp was used to draw the Bland-Altman plots and to determine the "conversion" equation given in Exhibit 46 (Carstensen et al., 2013).

T	G	T	G	T	G
0.96	0.85	2.28	2.17	3.19	3.15
1.16	1.00	2.15	2.20	3.12	3.15
0.97	1.00	2.29	2.28	3.33	3.40
1.01	1.00	2.45	2.43	3.51	3.42
1.25	1.20	2.40	2.55	3.66	3.62
1.22	1.20	2.79	2.60	3.95	3.95
1.46	1.38	2.77	2.65	4.20	4.27
1.66	1.65	2.64	2.67	4.05	4.30
1.75	1.68	2.73	2.70	4.30	4.35
1.72	1.70	2.67	2.70	4.74	4.75
1.67	1.70	2.61	2.70	4.71	4.79
1.67	1.70	3.01	3.00	4.71	4.80
1.93	1.88	2.93	3.02	4.74	4.80
1.99	2.00	3.18	3.03	5.23	5.42
2.01	2.05	3.18	3.11	6.21	6.20

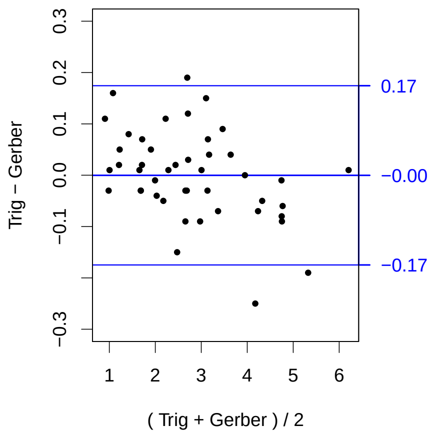


Exhibit 45: LEFT PANEL: Values of fat concentration in human milk (expressed in centigram per milliliter) determined by measurement of glycerol released by enzymatic hydrolysis of triglycerides (T) and by the Gerber method (G) (Lucas et al., 1987), as reported by Bland and Altman (1999, Table 3). RIGHT PANEL: Bland-Altman plot, with the average difference and the *limits of agreement* indicated by horizontal (blue) lines.

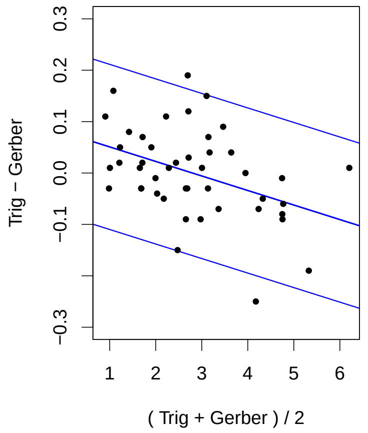


Exhibit 46: Bland-Altman plot recognizing that the differences between paired measured values depend on the averages of the same values. The corresponding equation that "converts" a value produced by the Gerber method into the value that Trig would be expected to produce is Trig $= 0.0779 + 0.9721 \times$ Gerber, with standard uncertainty 0.0792 cg/mL. The slope is consistent with the fact that only about 98 % of the fat in human milk is present as triglycerides (Lucas et al., 1987), which are the target of Trig.

E32 Load Cell Calibration. Calibrating a force transducer consists of characterizing its response R to different standard values of the applied force F. The data listed in Exhibit 47 are representative of a modern force transducer designed with extremely good control over sensitivities to force application alignment, sequencing, and timing. The response originates in a strain gage bridge network within the force transducer, and represents the ratio of the bridge output voltage (millivolt) to the bridge excitation voltage (volt).

NIST usually characterizes the transducer response by developing a calibration function ψ that, given a value of F, produces $R = \psi(F)$. To use the transducer to measure forces in practice, a function φ is needed that does the reverse: given an instrumental indication R, it produces an estimate of the applied force $F = \varphi(R)$. The traditional procedure (Bartel, 2005) has been to choose ψ as a polynomial of low degree, and to determine its coefficients by fitting the polynomial to values of R for given values of F by ordinary least squares. Subsequently, φ is defined as the mathematical inverse of ψ.

The traditional procedure ignores the uncertainty associated with the values of the applied

RUN 1		RUN 2		RUN 3	
F (kN)	R (mV/V)	F (kN)	R (mV/V)	F (kN)	R (mV/V)
222.411	0.088905	222.411	0.088906	222.411	0.088889
444.822	0.177777	444.822	0.177776	444.822	0.177767
889.644	0.355454	889.644	0.355453	889.644	0.355437
1334.467	0.533199	1334.467	0.533198	1334.467	0.533184
1779.289	0.710934	1779.289	0.710932	1779.289	0.710917
2224.111	0.888732	2224.111	0.888728	2224.111	0.888714
2668.933	1.066526	2668.933	1.066533	2668.933	1.066495
3113.755	1.244343	3113.755	1.244336	3113.755	1.244311
3558.578	1.422163	3558.578	1.422169	3558.578	1.422137
4003.400	1.600025	4003.400	1.600020	4003.400	1.599981
4448.222	1.777899	4448.222	1.777906	4448.222	1.777849

Exhibit 47: Forces and corresponding instrumental responses in three calibration runs of a load cell under compression. The forces are exactly the same in the three runs because they result from the application of the same dead-weights, and temporal differences in buoyancy are neglected.

forces when it determines the coefficients of the calibration function, but it does take them into account, as well as the uncertainty associated with the transducer response, when evaluating the calibration uncertainty. (Bartel, 2005) describes a most meticulous evaluation of the uncertainties associated with the forces and with the responses in calibrations performed in the NIST force laboratory, where the relative standard uncertainties associated with the forces applied during calibration, and with the electrical calibration of the voltage-ratio measuring instruments, both are 0.0005 %.

In this example, the function φ that is used to estimate the value of the applied force responsible for a particular transducer response is determined directly (and not as the inverse of the calibration function), by application of a version of the so-called *errors-in-variables* (EIV) model (Carroll et al., 2006).

The measurement model involves the following system of simultaneous observation equations: $R_{ij} = \rho_i + \delta_i + \omega_{ij}$, and $F_i = \varphi(\rho_i) + \varepsilon_i$, for $i = 1, \dots, m$ and for $j = 1, \dots, n$, where $m = 11$ is the number of force standards used during calibration, and $n = 3$ is the number of calibration runs (with each run involving the application of the same m standard forces). R_{ij}, with true value ρ_i, denotes the instrumental response read in the jth application of force F_i, whose true value is $\varphi(\rho_i)$.

In this case, the function φ is approximated by a polynomial of the second degree, which has been chosen from among polynomials of the first, second, and third degrees based on values of the Bayesian Information Criterion (BIC) (Burnham and Anderson, 2002), and on examination of plots of residuals.

The errors $\{\varepsilon_i\}$ are the differences between the true forces applied by the standard weights, and the values calculated for them based on the masses and volumes of the weights, and on the local gravitational acceleration and its vertical gradient. They are assumed to remain constant in the n calibration runs, and amount to 0.0005 % of the values of the true forces, $u(\varepsilon_i) = 0.000005\varphi(\rho_i)$. (NIST is currently developing a calibration procedure that takes into account changes in buoyancy effects attributable to changes in atmospheric conditions in the laboratory, which are measured in real-time, hence the applied forces are no longer assumed to remain invariant from run to run.)

The errors $\{\delta_i\}$ are the differences between the actual values of the instrumental responses and their true values that are attributable to uncertainty associated with the measurements of electrical quantities, including electrical calibration uncertainty. The $\{\delta_i\}$, too, are assumed to remain constant from run to run. The corresponding standard uncertainties amount to 0.0005 % of the values of the true responses, $u(\delta_i) = 0.000005\rho_i$.

The errors $\{\omega_{ij}\}$ describe the differences of the instrumental responses that are observed in different runs, which are attributable in large part to deliberate, between-run changes in the orientation of the transducer relative to the apparatus that applies forces to it. The errors that pertain to force F_i are assumed to have a common probability distribution with standard deviation σ_i, which is evaluated statistically (Type A evaluation). In this example, the same relative uncertainty was chosen for all the $\{\omega_{ij}\}$ as the median of the relative uncertainties $\{\sigma_i/\overline{R}_i\}$ where $\overline{R}_i$ denotes the average transducer response to force F_i over the n runs, for $i = 1, \dots, m$. It so turns out that the σ_i are about 3 times larger than the corresponding $u(\delta_i)$.

The model just described was fitted to the data in Exhibit 47 by the method of maximum likelihood (which in fact reduces to non-linear, weighted least squares), assuming that the $\{\varepsilon_i\}$, the $\{\delta_i\}$, and the $\{\omega_{ij}\}$ are like outcomes of independent, Gaussian random variables all with mean zero and with standard deviations equal to the standard uncertainties specified above. The corresponding optimization was done numerically using R (R Core Team, 2015) function `nloptr`, defined in the package of the same name (Ypma, 2014; Johnson, 2015), employing the the "Subplex" algorithm (Rowan, 1990). The fitting procedure produces estimates of the (3) coefficients of the second degree polynomial φ, and of the true values $\rho_1, \dots, \rho_m$ of the transducer responses.

The parametric statistical bootstrap was used for uncertainty evaluation, and it took into account the fairly small number of degrees of freedom that the relative standard uncertainty associated with the between-run dispersion of values is based on, and it involved repeating the following steps for $k = 1, \dots, K = 10\,000$:

1. Compute perturbed values of the applied forces as $F_{i,k} = \widehat{\varphi}(\widehat{\rho}_i) + \delta_{i,k}$, where $\widehat{\varphi}$ denotes the function fitted to the calibration data using the errors-in-variables procedure, $\widehat{\rho}_i$ denotes the estimate of the true instrumental response, and $\delta_{i,k}$ denotes a simulated error with mean 0 and standard deviation equal to $u(\delta_i)$, for $i = 1, \dots, m$.

2. Compute perturbed values of the transducer responses as $R_{ij,k} = \widehat{\rho}_i + \varepsilon_{i,k} + \omega_{ij,k}$, where $\varepsilon_{i,k}$ denotes a simulated error with mean 0 and standard deviation equal to $u(\varepsilon_i)$, and $\omega_{ij,k}$ denotes a simulated error with mean 0 and standard deviation equal to σ_i, for $i = 1, \dots, m$ and $j = 1, \dots, n$.

3. Compute the errors-in-variables estimate φ^*_k of the function that produces values of force given values of the transducer response.

All the simulated errors mentioned above are drawn from Gaussian distributions, except the $\{\omega_{ij,k}\}$, which are drawn from Student's t distributions with $m(n-1)$ degrees of freedom, rescaled to have the correct standard deviations. Exhibit 48 shows a coverage region (depicted as a shaded band) for the whole curve φ, computed by applying R function `envelope`, defined in package `boot` (Canty and Ripley, 2013a; Davison and Hinkley, 1997), to $\varphi^*_1, \dots, \varphi^*_K$.

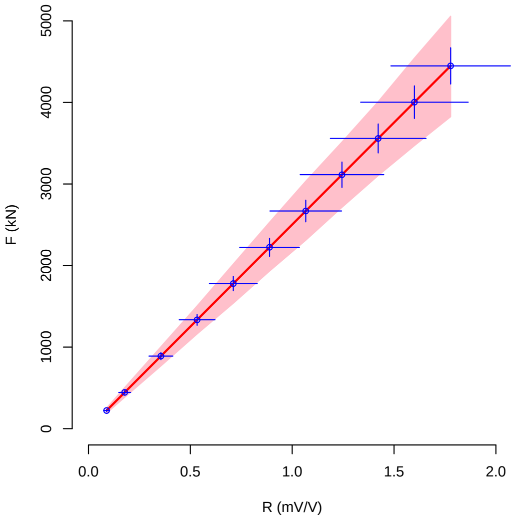


Exhibit 48: EIV regression function (red line) used to predict force as a function of instrumental response, calibration values (open blue circles) as listed in Exhibit 47, and associated standard uncertainties (horizontal and vertical segments), and approximate 95 % simultaneous coverage region (pink) for the regression function. Both the lengths of the segments representing the standard uncertainties, and the (vertical) thickness of the coverage band, are magnified 10 000 times. The relative standard uncertainties associated with the forces are all approximately equal to 0.0005 %.

E33 Atomic Weight of Hydrogen. Hydrogen has two stable isotopes, ^{1}H and ^{2}H, the former being far more abundant in normal materials than the latter, which is also known as deuterium. The Commission on Isotopic Abundances and Atomic Weights (CIAAW) of the International Union of Pure and Applied Chemistry (IUPAC), defines "normal material" for a particular element any terrestrial material that "is a reasonably possible source for this element or its compounds in commerce, for industry or science; the material is not itself studied for some extraordinary anomaly and its isotopic composition has not been modified significantly in a geologically brief period" (Peiser et al., 1984).

The atomic weight of hydrogen in a material is a weighted average of the masses of these isotopes, $m_a(^1\text{H}) = 1.007\,825\,032\,2$ Da and $m_a(^2\text{H}) = 2.014\,101\,778\,1$ Da, with weights proportional to the amount fractions of ^{1}H and ^{2}H in the material. Since these fractions vary between materials, the atomic weight of hydrogen (and of other elements that have more than one stable isotope) is not a constant of nature (Coplen and Holden, 2011). The standard uncertainties associated with those masses are $u(m_a(^1\text{H})) = u(m_a(^2\text{H})) = 0.000\,000\,000\,3$ Da (`www.ciaaw.org/hydrogen.htm`).)

Chesson et al. (2010, Table 2) reports $\delta^2\text{H} = 16.2$ ‰ measured by isotope ratio mass spectrometry in water extracted from a sample of Florida orange juice, and $\delta^2\text{H} = -16.8$ ‰ measured in a sample of Florida tap water. The corresponding standard measurement uncertainty was $u(\delta^2\text{H}) = 1.7$ ‰ (Lesley Chesson, 2015, personal communication).

Delta values (Coplen, 2011) express relative differences of isotope ratios in a sample and in a reference material, which for hydrogen is the Vienna Standard Mean Ocean Water (VSMOW) maintained by the International Atomic Energy Agency (Martin and Gröning, 2009) For example, $\delta^2\text{H} = (R(^2\text{H}/^1\text{H})_\text{M} - R(^2\text{H}/^1\text{H})_\text{VSMOW})/R(^2\text{H}/^1\text{H})_\text{VSMOW}$, where $R(^2\text{H}/^1\text{H})_\text{M}$ denotes the ratio of the numbers of atoms of ^{2}H and of ^{1}H in material M and $R(^2\text{H}/^1\text{H})_\text{VSMOW}$ $= 155.76 \times 10^{-6}$ (Wise and Watters, 2005a) is its counterpart for VSMOW.

Coplen et al. (2002, Page 1992) point out that "citrus trees in subtropical climates may undergo extensive evaporation, resulting in ^{2}H enrichment in cellular water". The question we wish to consider is whether the isotopic fractionation that led to the foregoing measured values of δ^2H is sufficient to substantiate a statistically significant difference between the atomic weight of hydrogen in the two materials, once the contributions from all relevant sources of uncertainty are taken into account.

The uncertainty for the atomic weight of hydrogen may be evaluated by application of the Monte Carlo method of the GUM-S1. Given a delta value $\delta^2\mathrm{H}_{\mathrm{M,VSMOW}}$ for a material M (either orange juice or tap water in this case), and the associated standard uncertainty $u(\delta^2\mathrm{H}_{\mathrm{M,VSMOW}})$, choose a suitably large sample size K, and repeat the following steps for $k = 1, \dots, K$:

1. Draw a value $\delta^2\mathrm{H}_{\mathrm{M,VSMOW},k}$ from a uniform (rectangular) distribution with mean $\delta^2\mathrm{H}_{\mathrm{M,VSMOW}}$ and standard deviation $u(\delta^2\mathrm{H}_{\mathrm{M,VSMOW}})$;
2. Draw a value $x_k(^2\mathrm{H})_{\mathrm{VSMOW}}$ for the amount fraction of ^{2}H in the VSMOW standard from a Gaussian distribution with mean $x(^2\mathrm{H})_{\mathrm{VSMOW}} = 0.999\,844\,26$ and standard deviation $u(x(^2\mathrm{H})_{\mathrm{VSMOW}}) = 0.000\,000\,025$;
3. Compute the corresponding amount fraction of ^{1}H in the standard, $x_k(^1\mathrm{H})_{\mathrm{VSMOW}} = 1 - x_k(^2\mathrm{H})_{\mathrm{VSMOW}}$;
4. Compute the isotope ratio in material M as $R_k(^2\mathrm{H}/^1\mathrm{H})_{\mathrm{M}} = (\delta^2\mathrm{H}_{\mathrm{M,VSMOW},k}+1)\, x_k(^2\mathrm{H})_{\mathrm{VSMOW}} / x_k(^1\mathrm{H})_{\mathrm{VSMOW}}$;
5. Compute the amount fraction of ^{2}H in material M as $x_k(^2\mathrm{H})_{\mathrm{M}} = R_k(^2\mathrm{H}/^1\mathrm{H})_{\mathrm{M}} / (1 + R_k(^2\mathrm{H}/^1\mathrm{H})_{\mathrm{M}})$;
6. The corresponding amount fraction of ^{1}H in material M is $x_k(^1\mathrm{H})_{\mathrm{M}} = 1 - x(^2\mathrm{H})_{\mathrm{M}}$;
7. Draw a value $m_{a,k}(^2\mathrm{H})$ from a uniform (rectangular) distribution with mean $m_a(^2\mathrm{H})$ and standard deviation $u(m_a(^2\mathrm{H}))$;
8. Draw a value $m_{a,k}(^1\mathrm{H})$ from a uniform (rectangular) distribution with mean $m_a(^1\mathrm{H})$ and standard deviation $u(m_a(^2\mathrm{H}))$;
9. Compute a sample value from the resulting probability distribution of the atomic weight of hydrogen in material M as $A_{\mathrm{r},k}(\mathrm{H})_{\mathrm{M}} = \left[x_k(^2\mathrm{H})_{\mathrm{M}} m_{\mathrm{a},k}(^1\mathrm{H}) + x_k(^1\mathrm{H})_{\mathrm{M}} m_{\mathrm{a},k}(^2\mathrm{H})\right] / m_{\mathrm{u}}$, where $m_{\mathrm{u}} = 1$ Da exactly.

These steps produce a sample of size K from the probability distribution of the atomic weight of hydrogen in material M that expresses uncertainty contributions from the following sources: measurement of the delta value, amount fractions of the two stable isotopes of hydrogen in the standard, and atomic masses of the two isotopes.

The mean of this sample of values of the atomic weight of hydrogen, $\{A_{\mathrm{r}}(\mathrm{B})_{\mathrm{M},1}, \dots, A_{\mathrm{r}}(\mathrm{B})_{\mathrm{M},K}\}$, is an estimate of the atomic weight of hydrogen in material M, and the standard deviation is an evaluation of the associated standard uncertainty $u(A_{\mathrm{r}}(\mathrm{H})_{\mathrm{M}})$.

For the measurements of the isotopic composition of orange juice (OJ) and tap water (TW), application of this procedure with $K = 1 \times 10^7$ produced 1.007 983 7 as estimate of $A_{\mathrm{r}}(\mathrm{H})_{\mathrm{OJ}}$, and 1.007 981 3 as estimate of $A_{\mathrm{r}}(\mathrm{H})_{\mathrm{TW}}$. The corresponding, associated standard uncertainties were both 0.000 000 3. Since $(1.0079837 - 1.0079813)/\sqrt{2 \times 0.0000003^2} = 5.7$, and the probability of a Gaussian random variable taking a value more than 5.7 standard deviations away from its mean is 2×10^{-8}, we conclude that the difference between the atomic weight of hydrogen in these samples of OJ and TW is statistically, highly significant.

E34 Atmospheric Carbon Dioxide. The concentration of CO_2 in the atmosphere has been measured regularly at Mauna Loa (Hawaii) since 1958 (Keeling et al., 1976). Etheridge et al. (1996) report values of the same concentration measured in air samples that became trapped in ice between 1006 and 1978, and that were recovered from several ice cores drilled in the region of the Law Dome (Antarctica). The yearly average concentrations from both locations (the series overlap between 1959 and 1978) are listed in Exhibit 49 and depicted graphically in Exhibit 50.

Law Dome						Mauna Loa			
yr	*c*	yr	*c*	yr	*c*	yr	*c*	yr	*c*
1006	279.4	1832	284.5	1939	309.2	1959	316.0	1987	349.2
1046	280.3	1840	283.0	1940	310.5	1960	316.9	1988	351.6
1096	282.4	1845	286.1	1944	309.7	1961	317.6	1989	353.1
1146	283.8	1850	285.2	1948	309.9	1962	318.4	1990	354.4
1196	283.9	1854	284.9	1948	311.4	1963	319.0	1991	355.6
1246	281.7	1861	286.6	1953	311.9	1964	319.6	1992	356.4
1327	283.4	1869	287.4	1953	311.0	1965	320.0	1993	357.1
1387	280.0	1877	288.8	1953	312.7	1966	321.4	1994	358.8
1387	280.4	1882	291.9	1954	313.6	1967	322.2	1995	360.8
1446	281.7	1886	293.7	1954	314.7	1968	323.0	1996	362.6
1465	279.6	1891	294.7	1954	314.1	1969	324.6	1997	363.7
1499	282.4	1892	294.6	1959	315.7	1970	325.7	1998	366.6
1527	283.2	1898	294.7	1962	318.7	1971	326.3	1999	368.3
1547	282.8	1899	296.5	1962	317.0	1972	327.4	2000	369.5
1570	281.9	1905	296.9	1962	319.4	1973	329.7	2001	371.1
1589	278.7	1905	298.5	1962	317.0	1974	330.2	2002	373.2
1604	274.3	1905	299.0	1963	318.2	1975	331.1	2003	375.8
1647	277.2	1912	300.7	1965	319.5	1976	332.1	2004	377.5
1679	275.9	1915	301.3	1965	318.8	1977	333.8	2005	379.8
1692	276.5	1924	304.8	1968	323.7	1978	335.4	2006	381.9
1720	277.5	1924	304.1	1969	323.2	1979	336.8	2007	383.8
1747	276.9	1926	305.0	1970	325.2	1980	338.7	2008	385.6
1749	277.2	1929	305.2	1970	324.7	1981	340.1	2009	387.4
1760	276.7	1932	307.8	1971	324.1	1982	341.4	2010	389.9
1777	279.5	1934	309.2	1973	328.1	1983	343.0	2011	391.6
1794	281.6	1936	307.9	1975	331.2	1984	344.6	2012	393.8
1796	283.7	1938	310.5	1978	335.2	1985	346.0	2013	396.5
1825	285.1	1939	311.0	1978	332.0	1986	347.4		

Exhibit 49: Yearly (yr) average amount-of-substance fraction *c* (expressed as micromole of CO_2 per mole of air) in the atmosphere, measured either in air bubbles trapped in ice at the Law Dome (Antarctica), or directly in the atmosphere at Mauna Loa (Hawaii).

The measurand is the function θ that produces the true value of the yearly average atmospheric concentration for any given year between 1006 and 2014. Since θ may be expected to vary smoothly over this range, estimating it amounts to building a smooth interpolant for the data, which can be done in many different ways.

The observation equation (statistical model) selected for illustration in this example is a treed Gaussian process (Gramacy and Lee, 2008), which makes no assumptions about the functional form of θ, and represents the target function either as an outcome of a single Gaussian random function, or as an outcome of two of more Gaussian random functions joined end-to-end.

A Gaussian random function is a collection of correlated Gaussian random variables $\{\theta(t) : t = 1006, \ldots, 2013\}$, also called a Gaussian *stochastic process*. The correlations allow the function to capture the fact that values at neighboring epochs tend to be more similar than values at widely separated epochs. Such functions can enjoy much greater modeling flexibility than a polynomial or even a piecewise polynomial function, for example.

The function btgp defined in R package tgp (Gramacy, 2007) implements a Bayesian procedure to fit this model (Gramacy and Lee, 2008; Chipman et al., 2013). When fitted to this data, a change in regime around 1877 was detected. Exhibit 50 shows the estimate of θ and a 95 % coverage band.

The thick tick mark pointing up from the horizontal axis indicates the year 1877, which marks the transition from a regimen that lasted for at least 800 years, during which the amount fraction of CO_2 remained fairly constant at about 280 µmol/mol, to a period that started with the Industrial Revolution and continues until the present, when this amount fraction has been increasing very rapidly.

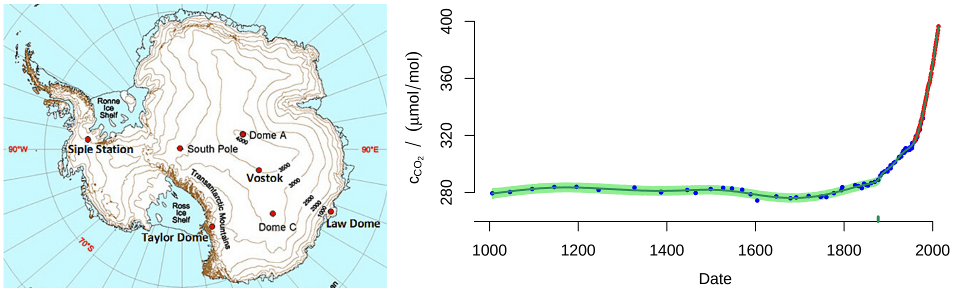


Exhibit 50: Antarctica ice-core locations, including the Law Dome (left panel: image source cdiac.ornl.gov/trends/co2/ice_core_co2.html, Carbon Dioxide Information Analysis Center). Yearly average amount fraction c_{CO_2} (right panel, micromole of CO_2 per mole of air) measured either in air bubbles trapped in ice at the Law Dome (large blue dots), or directly in the atmosphere at Mauna Loa, Hawaii (small red dots). Estimate of the function θ that produces the true value of c_{CO_2} for any given year between 1006 and 2014 (solid, dark green line), qualified with a 95 % coverage band. The green tick mark pointing up from the horizontal axis indicates the year (1877) when a transition occurs in the model, from one Gaussian process to another.

E35 Colorado Uranium. The National Geochemical Survey maintained by the U.S. Geological Survey (U.S. Geological Survey Open-File Report 2004-1001, version 5.0 available at mrdata.usgs.gov/geochem/doc/home.htm, accessed on March 22, 2015) includes data for 1150 samples, primarily of stream sediments, collected in Colorado between 1975 and 1980 as part of the National Uranium Resource Evaluation (NURE) program (Smith, 2001). The corresponding data may be downloaded (in any one of several formats) from mrdata.usgs.gov/geochem/select.php?place=fUS08&div=fips.

The mass fraction of uranium in these samples was measured using delayed neutron counting (Knight and McKown, 2002). The measured values range from 1 mg/kg to 147 mg/kg, and their distribution is markedly asymmetric, with right tail much longer than the left. A Box-Cox transformation that re-expresses an observed value x into $(x^{\lambda} - 1)/\lambda$, with $\lambda = -0.7$, reduces such asymmetry substantially (Box and Cox, 1964), and enhances the plausibility of models that involve Gaussian assumptions.

The measurand is the function θ that, given the geographical coordinates (u, v) of a location within Colorado, produces an estimate of the mass fraction of uranium in sediments at that location. The generic observation equation expresses the measured value of the mass fraction of uranium $w(u,v)$ as $(w(u,v)^{\lambda} - 1)/\lambda = \theta(u,v) + \varepsilon(u,v)$.

The measurement errors $\{\varepsilon(u,v)\}$ are assumed to behave like values of independent, Gaussian random variables with mean zero and the same standard deviation. The function θ is deterministic in two of the models considered below, and stochastic in two others.

A polynomial (in the geographical coordinates) would be an example of a deterministic function. A collection of correlated Gaussian random variables, where each one describes the mass fraction of uranium at one location in the region, would be an example of a stochastic function: the correlations capture the fact that neighboring locations tend to have more similar values of that mass fraction than locations that are far apart.

Exhibit 51 shows that both deterministic and stochastic functions are able to capture very much the same patterns in the spatial variability of the data. Even though the function θ can be evaluated at any location throughout the region, here it is displayed as an image that depicts the values that θ takes at the center of each pixel in a regular grid comprising 40×30 pixels. These are the four models used for θ:

Q: Locally quadratic regression model with nearest-neighbor component of the smoothing parameter chosen by cross-validation, as implemented in R package `locfit` (Loader, 1999, 2013);

K: Ordinary kriging model with Matérn's covariance function and estimation of spatial anisotropy as implemented in R package `intamap` (Stein, 1999; Pebesma et al., 2010);

G: Generalized additive model with thin plate regression splines and smoothing parameter chosen by generalized cross-validation, as implemented in R package `mgcv` (Wood, 2003, 2006);

L: Multi-resolution Gaussian process model as implemented in R package `LatticeKrig`, with default settings for all the user adjustable parameters (Nychka et al., 2013, 2014).

The four estimates of θ are generally similar but clearly differ in many details. The significance of these differences depends on the uncertainty associated with each estimate. Instead of exploring the differences, we may choose instead to combine the estimates, and then to capture the differences that are attributable to model uncertainty alongside other identifiable sources of uncertainty, when evaluating the uncertainty associated with the result.

Model averaging is often used for this purpose (Hoeting et al., 1999; Clyde and George, 2004), which typically is done by computing the weighted mean of the results corresponding to the different models, with weights proportional to the Bayesian posterior probabilities of the models given the data.

In this case, we adopt the simplest version possible of model averaging, which assigns to the pixel with center coordinates (u,v) the (unweighted) arithmetic average of the values that the four estimates described above take at this location: $\widehat{\theta}(u,v) = (\widehat{\theta}_{\mathrm{Q}}(u,v) + \widehat{\theta}_{\mathrm{K}}(u,v) + \widehat{\theta}_{\mathrm{G}}(u,v) + \widehat{\theta}_{\mathrm{L}}(u,v))/4$.

Exhibit 52 shows this pixelwise average, and also the endpoints of pixelwise coverage intervals for θ (one interval for each of the pixels in the image) based on a Monte Carlo sample of size $K = 1000$ drawn from the probability distribution of θ. Each element in this sample,

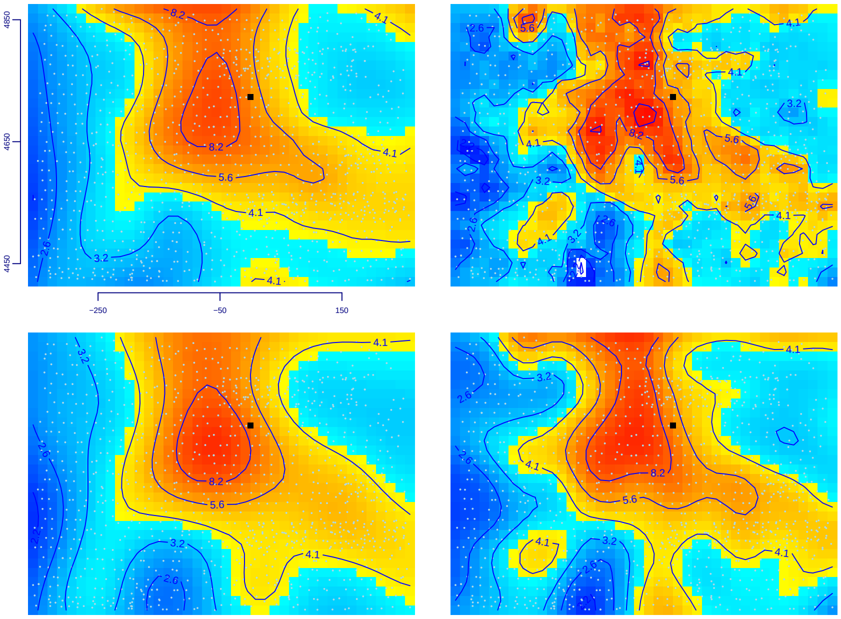


Exhibit 51: Four estimates of the spatial distribution of the mass fraction of uranium in stream sediments throughout Colorado: (i) **Q** (locally quadratic regression, top left); (ii) **K** (ordinary kriging, top right); (iii) **G** (generalized additive model, bottom left); and (iv) **L** (multi-resolution Gaussian process model, bottom right). The black square marks the location of the city of Denver, and the light-colored little dots mark the locations that were sampled. The geographical coordinates are expressed in km, and the labels of the contour lines are expressed in mg/kg.

for $k = 1, \dots, K$, is a map built as follows, where $m = 1150$ denotes the number of locations where the mass fraction of uranium was measured:

1. Draw a sample of size m, uniformly at random and with replacement, from the set of m locations where sediment was collected for analysis;
2. Since the same location may be selected more than once, the geographical coordinates of all the locations that are drawn into the sample are jittered slightly, to avoid the occurrence of duplicated locations, which some of the software used cannot handle;
3. Obtain estimates $\widehat{\theta}_{\mathrm{Q},k}$, $\widehat{\theta}_{\mathrm{K},k}$, $\widehat{\theta}_{\mathrm{G},k}$, $\widehat{\theta}_{\mathrm{L},k}$ as described above, but using the sample drawn from the original data;
4. Compute $\widehat{\theta}^*_k = (\widehat{\theta}_{\mathrm{Q},k} + \widehat{\theta}_{\mathrm{K},k} + \widehat{\theta}_{\mathrm{G},k} + \widehat{\theta}_{\mathrm{L},k})/4$.

The coverage interval at the pixel whose center has coordinates (u, v) has left and right endpoints equal to the 2.5th and 97.5th percentiles of $\{\widehat{\theta}^*_1(u, v), \dots, \widehat{\theta}^*_K(u, v)\}$. These maps of percentiles indicate how much, or how little of the structures apparent in $\widehat{\theta}$ are significant once measurement uncertainty (including model uncertainty) is taken into account.

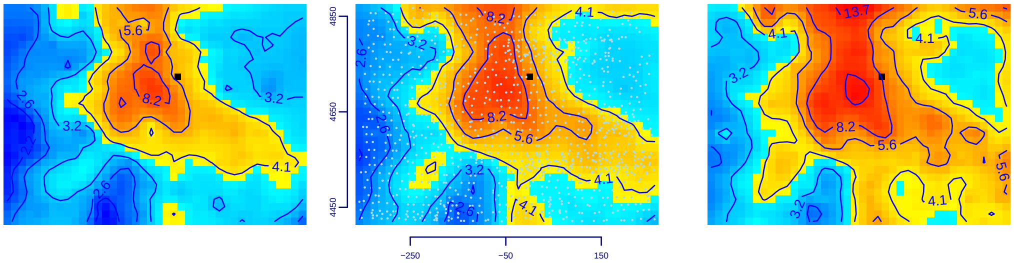


Exhibit 52: The center panel shows the pointwise average of the four estimates of the spatial distribution of the mass fraction of uranium depicted in Exhibit 51. The black dot marks the location of the city of Denver, and the light-colored little dots mark the locations that were sampled. The geographical coordinates are expressed in km, and the labels of the contour lines are expressed in mg/kg. The left and right panels show the left and right end-points of approximate 95 % coverage intervals for θ.

References

R. J. Adcock. A problem in least squares. *The Analyst*, 5:53–54, March 1878.

Agilent. *Fundamentals of RF and Microwave Power Measurements (Part 3) — Power Measurement Uncertainty per International Guides*. Agilent Technologies, Santa Clara, CA, April 2011. URL `http://cp.literature.agilent.com/litweb/pdf/5988-9215EN.pdf`. Application Note 1449-3, Literature Number 5988-9215EN.

D. G. Altman and J. M. Bland. Measurement in medicine: the analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician),*, 32(3):307–317, September 1983.

Analytical Methods Committee. Robust Statistics — How Not to Reject Outliers. Part 1. Basic Concepts. *Analyst*, 114:1693–1697, December 1989. doi: 10.1039/AN9891401693. Royal Society of Chemistry.

T. W. Anderson and D. A. Darling. Asymptotic theory of certain “goodness-of-fit” criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1952.

ANSI/NCSL. *Requirements for the Calibration of Measuring and Test Equipment*. American National Standards Institute / National Conference of Standards Laboratories International, Washington, DC / Boulder, CO, March 2013. ANSI/NCSL Z540.3-2006 (R2013).

T. J. Aragón. *epitools: Epidemiology Tools*, 2012. URL `http://CRAN.R-project.org/package=epitools`. R package version 0.5-7.

R. A. Askey and R. Roy. Gamma function. In F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors, *NIST Handbook of Mathematical Functions*. Cambridge University Press, Cambridge, UK, 2010.

G. Audi, F.G. Kondev, M. Wang, B. Pfeiffer, X. Sun, J. Blachot, and M. MacCormick. The NUBASE2012 evaluation of nuclear properties. *Chinese Physics C*, 36(12):1157–1286, December 2012.

T. Aven, E. Zio, P Baraldi, and R. Flage. *Uncertainty in Risk Assessment: The Representation and Treatment of Uncertainties by Probabilistic and Non-Probabilistic Methods*. John Wiley & Sons, Chichester, UK, 2014.

K. Bache and M. Lichman. UCI Machine Learning Repository, 2013. URL `http://archive.ics.uci.edu/ml`.

M. C. Baddeley, A. Curtis, and R. Wood. An introduction to prior information derived from probabilistic judgements: elicitation of knowledge, cognitive bias and herding. In A. Curtis and R. Wood, editors, *Geological Prior Information: Informing Science and Engineering*, volume 239 of *Special Publications*, pages 15–27. Geological Society, London, 2004. doi: 10.1144/GSL.SP.2004.239.01.02.

R. Badertscher, T. Berger, and R. Kuhn. Densitometric determination of the fat content of milk and milk products. *International Dairy Journal*, 17(1):20–23, 2007. doi: 10.1016/j.idairyj.2005.12.013.

E. Barkan and B. Luz. The relationships among the three stable isotopes of oxygen in air, seawater and marine photosynthesis. *Rapid Communications in Mass Spectrometry*, 25(16):2367–2369, 2011. doi: 10.1002/rcm.5125. Letter to the Editor.

R. Barlow. Asymmetric errors. In *PHYSTAT2003: Statistical Problems in Particle Physics, Astrophysics and Cosmology*, pages 250–255, Menlo Park, CA, September 8th–11th 2003. SLAC National Accelerator Laboratory. URL `http://www.slac.stanford.edu/econf/C030908/proceedings.html`.

T. Bartel. Uncertainty in NIST force measurements. *Journal of Research of the National Institute of Standards and Technology*, 110(6):589–603, 2005.

S. Bell. *A Beginner's Guide to Uncertainty of Measurement*. Number 11 (Issue 2) in Measurement Good Practice Guide. National Physical Laboratory, Teddington, Middlesex, United Kingdom, 1999. URL `http://www.npl.co.uk/publications/guides/a-beginners-guide-to-uncertainty-of-measurement`. Amendments March 2001.

T. Benaglia, D. Chauveau, D. R. Hunter, and D. Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009. URL `http://www.jstatsoft.org/v32/i06/`.

J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley & Sons, Chichester, England, 2nd edition, 2007.

B. J. Biggerstaff and D. Jackson. The exact distribution of cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine*, 27:6093–6110, 2008. doi: 10.1002/sim.3428.

B. J. Biggerstaff and R. L. Tweedie. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, 16:753–768, 1997.

BIPM. *The International System of Units (SI)*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 8th edition, 2006.

J. M. Bland and D. G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327:307–310, 1986. doi:10.1016/S0140-6736(86)90837-8.

J. M. Bland and D. G. Altman. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8:135–160, 1999.

B. Bolker and R Development Core Team. *bbmle: Tools for general maximum likelihood estimation*, 2014. URL `http://CRAN.R-project.org/package=bbmle`. R package version 1.0.17.

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

W. A. Brand, S. S. Assonov, and T. B. Coplen. Correction for the ^{17}O interference in $\delta(^{13}\mathrm{C})$ measurements when analyzing CO_2 with stable isotope mass spectrometry (IUPAC Technical Report). *Pure and Applied Chemistry*, 82(8):1719–1733, 2010. doi: 10.1351/PAC-REP-09-01-05.

P. W. Bridgman. *The logic of modern physics*. The Macmillan Co., New York, NY, 1927.

S. E. Brockwell and I. R. Gordon. A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, 20:825–840, 2001.

K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, NY, 2nd edition, 2002.

T. A. Butler and J. L. Molloy. Preparation of candidate srm1641e mercury in water. Report of Analysis 646.01-14-025, National Institute of Standards and Technology, Material Measurement Laboratory, Chemical Sciences Division, Gaithersburg, MD, June 2014.

A. Canty and B. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2013a. URL `http://cran.r-project.org/web/packages/boot/`. R package version 1.3-9.

A. Canty and B. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2013b. URL `http://cran.r-project.org/web/packages/boot/`. R package version 1.3-15.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models — A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, Florida, second edition, 2006.

B. Carstensen. *Comparing Clinical Measurement Methods*. John Wiley & Sons, Chichester, UK, 2010.

B. Carstensen, L. Gurrin, C. Ekstrom, and M. Figurski. *MethComp: Functions for analysis of agreement in method comparison studies*, 2013. URL `http://CRAN.R-project.org/package=MethComp`. R package version 1.22.

H. Cavendish. Experiments to determine the density of the earth. by Henry Cavendish, Esq. F. R. S. and A. S. *Philosophical Transactions of the Royal Society of London*, 88:469–526, 1798. doi: 10.1098/rstl.1798.0022.

CDF Collaboration and D0 Collaboration. Combination of cdf and d0 w-boson mass measurements. *Physical Review D*, 88:052018, September 2013. doi: 10.1103/PhysRevD.88.052018.

J. M. Chambers. Linear models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, chapter 4. Chapman and Hall/CRC, Boca Raton, FL, 1991.

L. A. Chesson, G. J. Bowen, and J. R. Ehleringer. Analysis of the hydrogen and oxygen stable isotope ratios of beverage waters without prior water extraction using isotope ratio infrared spectroscopy. *Rapid Communications in Mass Spectrometry*, 24(21):3205–3213, 2010. doi: 10.1002/rcm.4759.

H. Chipman, E. I. George, R. B. Gramacy, and R. McCulloch. Bayesian treed response surface models. *Wiley*

Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(4):298–305, 2013. doi: 10.1002/widm.1094.

M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19:81–94, 2004.

H. Cooper, L. V. Hedges, and J. C. Valentine, editors. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation Publications, New York, NY, 2nd edition, 2009.

T. B. Coplen. Guidelines and recommended terms for expression of stable-isotope-ratio and gas-ratio measurement results. *Rapid Communications in Mass Spectrometry*, 25(17):2538–2560, 2011. doi: 10.1002/rcm.5129.

T. B. Coplen and N. E. Holden. Atomic weights: No longer constants of nature. *Chemistry International*, 33 (2), March-April 2011.

T. B. Coplen, J. K. Böhlke, P. De Bièvre, T. Ding, N. E. Holden, J. A. Hopple, H. R. Krouse, A. Lamberty, H. S. Peiser, K. Révész, S. E. Rieder, K. J. R. Rosman, E. Roth, P. D. P. Taylor, Jr. R. D. Vocke, and Y. K. Xiao. Isotope-abundance variations of selected elements. *Pure and Applied Chemistry*, 74(10):1987–2017, 2002.

A. Curtis and R. Wood. Optimal elicitation of probabilistic information from experts. In A. Curtis and R. Wood, editors, *Geological Prior Information: Informing Science and Engineering*, volume 239 of *Special Publications*, pages 127–145. Geological Society, London, 2004. doi: 10.1144/GSL.SP.2004.239.01.02.

F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, March 1964. doi: 10.1145/363958.363994.

S. Davidson, M. Perkin, and M. Buckley. *The Measurement of Mass and Weight*. National Physical Laboratory (NPL), Teddington, United Kingdom, June 2004. Measurement Good Practice Guide No. 71.

A. C. Davison and D. Hinkley. *Bootstrap Methods and their Applications*. Cambridge University Press, New York, NY, 1997. URL `http://statwww.epfl.ch/davison/BMA/`.

M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison-Wesley, 4th edition, 2011.

W. E. Deming. *Statistical Adjustment of Data*. John Wiley & Sons, New York, NY, 1943.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, September 1986.

T. Doiron and J. Beers. The Gauge Block Handbook. NIST Monograph 180, National Institute of Standards and Technology, Gaithersburg, MD, June 1995. URL `http://emtoolbox.nist.gov/Publications/NISTMonograph180.asp`. Corrections 2005.

T. Doiron and J. Stoup. Uncertainty and dimensional calibrations. *Journal of Research of the National Institute of Standards and Technology*, 102(6):647–676, November-December 1997.

A. M. Dziewonski and D. L. Anderson. Preliminary reference earth model. *Physics of the Earth and Planetary Interiors*, 25(4):297–356, 1981. doi: 10.1016/0031-9201(81)90046-7.

EA Laboratory Committee. *Expression of the Uncertainty of Measurement in Calibration*. EA-4/02 M: 2013. European co-operation for Accreditation, Paris, France, September 2013. Rev. 01.

B. Edlén. The refractive index of air. *Metrologia*, 2(2):71–80, 1966.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, UK, 1993.

C. Ehrlich. Terminological aspects of the *Guide to the Expression of Uncertainty in Measurement* (gum). *Metrologia*, 51(4):S145–S154, August 2014. doi: 10.1088/0026-1394/51/4/S145.

S. L. R. Ellison and A. Williams, editors. *Quantifying Uncertainty in Analytical Measurement*. EURACHEM/CITAC Guide CG-4, QUAM:2012.P1. EURACHEM, Third edition, 2012. URL `http://www.eurachem.org`.

D. M. Etheridge, L. P. Steele, R. L. Langenfelds, R. J. Francey, J.-M. Barnola, and V. I. Morgan. Natural and anthropogenic changes in atmospheric CO_2 over the last 1000 years from air in Antarctic ice and firn. *Journal of Geophysical Research*, 101:4115–4128, 1996. doi: 10.1029/95JD03410.

European Food Safety Authority. Guidance on the use of probabilistic methodology for modelling dietary exposure to pesticide residues. *EFSA Journal*, 10:2839 (95 pp.), 2012. doi: 10.2903/j.efsa.2012.2839. Scientific Opinion — EFSA Panel on Plant Protection Products and their Residues (PPR).

I. W. Evett and E. J. Spiehler. Rule induction in forensic science. In *KBS in Goverment*, pages 107–118, Pinner, UK, 1987. Online Publications.

C.-J. L. Farrell, S. Martin, B. McWhinney, I. Straub, P. Williams, and M. Herrmann. State-of-the-art vitamin D assays: A comparison of automated immunoassays with liquid chromatography-tandem mass spectrometry methods. *Clinical Chemistry*, 58(3):531–542, 2012. doi: 10.1373/clinchem.2011.172155.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons, New York, 3rd edition, 1968. Revised Printing.

R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10:507–521, May 1915. doi: 10.2307/2331838.

R. A. Fisher. On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, 1: 3–32, 1921.

J. Fox and S. Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011. URL `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`.

G. T. Fraser and R. L. Watters. *Standard Reference Material 1822a, Refractive Index Standard*. Office of Standard Reference Materials, National Institute of Standards and Technology, Department of Commerce, Gaithersburg, Maryland, 2008. URL `http://www.nist.gov/srm/`.

D. Freedman, R. Pisani, and R. Purves. *Statistics*. W. W. Norton & Company, New York, NY, fourth edition, 2007.

D. A. Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, New York, NY, 2009.

X. Fuentes-Arderiu and D. Dot-Bach. Measurement uncertainty in manual differential leukocyte counting. *Clinical Chemistry and Laboratory Medicine*, 47(1):112–115, 2009. doi: 10.1515/CCLM.2009.014.

X. Fuentes-Arderiu, M. García-Panyella, and D. Dot-Bach. Between-examiner reproducibility in manual differential leukocyte counting. *Accreditation and Quality Assurance*, 12:643–645, 2007. doi: 10.1007/s00769-007-0323-0.

C. Gauss. Theoria combinationis observationum erroribus minimis obnoxiae. In *Werke, Band IV, Wahrscheinlichkeitsrechnung und Geometrie*. Könighlichen Gesellschaft der Wissenschaften, Göttingen, 1823. http://gdz.sub.uni-goettingen.de/.

D. P. Gaver, D. Draper, P. K. Goel, J. B. Greenhouse, L. V. Hedges, C. N. Morris, and C. Waternaux. *Combining information: Statistical Issues and Opportunities for Research*. National Academy Press, Washington, DC, 1992. Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences, Commission on Physical Sciences, Mathematics and Applications, National Research Council.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall / CRC, Boca Raton, FL, 3rd edition, 2013.

C. J. Geyer and L. T. Johnson. *mcmc: Markov Chain Monte Carlo*, 2014. URL `http://CRAN.R-project.org/package=mcmc`. R package version 0.9-3.

A. Giordani and L. Mari. Measurement, models, and uncertainty. *IEEE TRansactions on Instrumentation and Measurement*, 61(8):2144–2152, August 2012.

R. B. Gramacy. tgp: An R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *Journal of Statistical Software*, 19(9):1–46, 2007. URL `http://www.jstatsoft.org/v19/i09`.

R. B. Gramacy and H. K. H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103:1119–1130, 2008.

GREAT Group. Feasibility, safety, and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial. *British Medical Journal*, 305(6853):548–553, 1992. doi: 10.1136/bmj.305.6853.548.

F. R. Guenther and A. Possolo. Calibration and uncertainty assessment for certified reference gas mixtures. *Analytical and Bioanalytical Chemistry*, 399:489–500, 2011.

R. G. Gullberg. Estimating the measurement uncertainty in forensic blood alcohol analysis. *Journal of Analytical Toxicology*, 36:153–161, 2012. doi: 10.1093/jat/bks012.

I. A. Harris and F. L. Warner. Re-examination of mismatch uncertainty when measuring microwave power and attenuation. *Microwaves, Optics and Antennas, IEE Proceedings H*, 128:35–41, February 1981.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, second edition, 2009. URL `http://statweb.stanford.edu/~tibs/ElemStatLearn/`.

T. Hastie, R. Tibshirani, F. Leisch, K. Hornik, and B. D. Ripley. *mda: Mixture and flexible discriminant analysis*, 2013. URL `http://CRAN.R-project.org/package=mda`. R package version 0.4-4.

L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, San Diego, CA, 1985.

J. P. T. Higgins and S. G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21: 1539–1558, 2002. doi: 10.1002/sim.1186.

J. P. T. Higgins, S. G. Thompson, and D. J. Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 172(1):137–159, January 2009.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*. John Wiley & Sons, New York, NY, 2nd edition, 1999.

P. Hosmer, H. Schatz, A. Aprahamian, O. Arndt, R. R. C. Clement, A. Estrade, K. Farouqi, K.-L. Kratz, S. N. Liddick, A. F. Lisetskiy, P. F. Mantica, P. Möller, W. F. Mueller, F. Montes, A. C. Morton, M. Ouellette, E. Pellegrini, J. Pereira, B. Pfeiffer, P. Reeder, P. Santi, M. Steiner, A. Stolz, B. E. Tomlin, W. B. Walters, and A. Wöhr. Half-lives and branchings for β-delayed neutron emission for neutron-rich Co-Cu isotopes in the r-process. *Physical Review C*, 82:025806, August 2010. doi: 10.1103/PhysRevC.82.025806.

F. Huber and C. Schmidt-Petri, editors. *Degrees of Belief*. Number 342 in Synthese Library. Springer Science+Business Media B.V., Dordrecht, Netherlands, 2009. ISBN 978-1-4020-9197-1.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2009.

ISO. *Gas analysis — Comparison methods for determining and checking the composition of calibration gas mixtures*. International Organization for Standardization (ISO), Geneva, Switzerland, 2001. International Standard ISO 6143:2001(E).

H. K. Iyer, C. M. J. Wang, and T. Mathew. Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, 99(468):1060–1071, December 2004.

H. Jeffreys. *Theory of Probability*. Oxford University Press, London, 3rd edition, 1961. Corrected Impression, 1967.

F. A. Jenkins and H. E. White. *Fundamentals of Optics*. McGraw-Hill, New York, New York, 4th edition, 1976.

N. P. Jewell. *Statistics for Epidemiology*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

N. L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1972. ISBN 0–471–44370–0.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, Volume 1*. John Wiley & Sons, New York, NY, Second edition, 1994.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, Volume 2*. John Wiley & Sons, New York, NY, Second edition, 1995.

N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate Discrete Distributions*. John Wiley & Sons, Hoboken, NJ, Third edition, 2005.

S. G. Johnson. The NLopt nonlinear-optimization package. `http://ab-initio.mit.edu/nlopt`, 2015. Last visited August 21, 2015.

Joint Committee for Guides in Metrology. *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008a. URL `http://www.bipm.org/en/publications/guides/gum.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections.

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, 2008b. URL `http://www.bipm.org/en/publications/guides/gum.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 101:2008.

Joint Committee for Guides in Metrology. *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008c. URL `http://www.bipm.org/en/publications/guides/vim.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 200:2008.

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 `http://www.bipm.org/en/publications/guides/gum.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 102:2011.

R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, September 1996.

C. D. Keeling, R. B. Bacastow, A. E. Bainbridge, C. A. Ekdahl, P. R. Guenther, and L. S. Waterman. Atmospheric carbon dioxide variations at Mauna Loa Observatory, Hawaii. *Tellus*, 28:538–551, 1976.

F. Killmann and E. von Collani. A note on the convolution of the uniform and related distributions and their use in quality control. *Economic Quality Control*, 16(1):17–41, 2001.

S. J. Kline and F. A. McClintock. Describing uncertainties in single-sample experiments. *Mechanical Engineering*, 75(1):3–8, 1953.

G. Knapp and J. Hartung. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22:2693–2710, 2003. doi: 10.1002/sim.1482.

R. J. Knight and D. M. McKown. Uranium and thorium by delayed neutron counting. In J. E. Taggart, editor, *Analytical methods for chemical analysis of geologic and othermaterials, U.S. Geological Survey*, pages Z1–Z5. U.S. Geological Survey, U.S. Department of the Interior, Denver, CO, 2002. Open-File Report 02-223.

S. Kotochigova, Z. H. Levine, E. L. Shirley, M. D. Stiles, and C. W. Clark. Local-density-functional calculations of the energy of atoms. *Physical Review A*, 55:191–199, January 1997a. doi: 10.1103/PhysRevA.55.191.

S. Kotochigova, Z. H. Levine, E. L. Shirley, M. D. Stiles, and C. W. Clark. Erratum: Local-density-functional calculations of the energy of atoms. *Physical Review A*, 56:5191–5192, December 1997b. doi: 10.1103/PhysRevA.56.5191.2.

S. Kotochigova, Z. H. Levine, E. L. Shirley, M. D. Stiles, and C. W. Clark. Atomic reference data for electronic structure calculations. National Institute of Standards and Technology, Gaithersburg, MD, December 2011. URL `http://www.nist.gov/pml/data/dftdata`. NIST Standard Reference Database 141.

A. Kramida, Y. Ralchenko, and J. Reader. Nist atomic spectra database. National Institute of Standards and Technology, Gaithersburg, MD, September 2013. URL `http://www.nist.gov/pml/data/asd.cfm`. NIST Standard Reference Database 78.

J. K. Kruschke. Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, 142 (2):573–603, 2013.

J. K. Kruschke and M. Meredith. *BEST: Bayesian Estimation Supersedes the t-Test*, 2013. URL `http://CRAN.R-project.org/package=BEST`. R package version 0.2.0.

T. Lafarge and A. Possolo. The NIST Uncertainty Machine. *NCLSI Measure Journal of Measurement Science*, 10(3), September 2015.

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics — Doklady*, 10(8):707–710, February 1966.

D. Lindley. *Understanding Uncertainty*. John Wiley & Sons, Hoboken, New Jersey, 2006.

D. V. Lindley. The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science*, 2(1):17–24, February 1987. doi: 10.1214/ss/1177013427.

T. P. J. Linsinger, J. Pauwels, A. Lamberty, H.G. Schimmel, A. M. H. van der Veen, and L. Siekmann. Estimating the uncertainty of stability for matrix crms. *Fresenius' Journal of Analytical Chemistry*, 370:183–188, 2001.

C. Loader. *Local Regression and Likelihood*. Springer-Verlag, New York, 1999.

C. Loader. *locfit: Local Regression, Likelihood and Density Estimation.*, 2013. URL `http://CRAN.R-project.org/package=locfit`. R package version 1.5-9.1.

A. Lucas, G. J. Hudson, P. Simpson, T. J. Cole, and B. A. Baker. An automated enzymic micromethod for the measurement of fat in human milk. *Journal of Dairy Research*, 54:487–492, November 1987. doi: 10.1017/S0022029900025693.

L. L. Lucas and M. P. Unterweger. Comprehensive review and critical evaluation of the half-life of tritium. *Journal of Research of the National Institute of Standards and Technology*, 105(4):541–549, July-August 2000.

R. D. Luce. The ongoing dialog between empirical science and measurement theory. *Journal of Mathematical Psychology*, 40:78–98, 1996.

A. Lymer. Calculating mismatch uncertainty. *Microwave Journal — Industry News*, 15 May 2008. URL `www.microwavejournal.com/articles/6166-calculating-mismatch-uncertainty`.

J. Mandel and R. Paule. Interlaboratory evaluation of a material with unequal numbers of replicates. *Analytical Chemistry*, 42(11):1194–1197, 1970. doi: 10.1021/ac60293a019.

J. Mandel and R. Paule. Correction — interlaboratory evaluation of a material with unequal numbers of replicates. *Analytical Chemistry*, 43(10):1287–1287, 1971. doi: 10.1021/ac60304a001.

E. Manuilova, A. Schuetzenmeister, and F. Model. *mcr: Method Comparison Regression*, 2014. URL `http://CRAN.R-project.org/package=mcr`. R package version 1.2.1.

L. Mari and P. Carbone. Measurement fundamentals: a pragmatic view. *IEEE Transactions on Instrumentation and Measurement*, 61(8):2107–2115, August 2012.

P. Martin and M. Gröning. *VSMOW2 & SLAP2 Reference Sheet for International Measurement Standards*. International Atomic Energy Agency, Vienna, Austria, May 2009.

Y. Matsuhisa, J. R. Goldsmith, and R. N. Clayton. Mechanisms of hydrothermal crystallization of quartz at 250 °C and 15 kbar. *Geochimica et Cosmochimica Acta*, 42(2):173–182, 1978. doi: 10.1016/0016-7037(78)90130-8.

D. Mauchant, K. D. Rice, M. A. Riley, D. Leber, D. Samarov, and A. L. Forster. Analysis of three different regression models to estimate the ballistic performance of new and environmentally conditioned body armor. Technical Report NISTIR 7760, National Institute of Standards and Technology, Gaithersburg, MD, February 2011.

H. A. J. Meijer and W. J. Li. The use of electrolysis for accurate δ^{17}O and δ^{18}O isotope measurements in water.

Isotopes in Environmental and Health Studies, 34(4):349–369, 1998. doi: 10.1080/10256019808234072.

R. G. Miller. *Simultaneous Statistical Inference*. Springer, New York, 2nd edition, 1981.

R. G. Miller. *Beyond ANOVA, Basics of Applied Statistics*. John Wiley & Sons, New York, NY, 1986.

M. Miyabe, C. Geppert, M. Kato, M. Oba, I. Wakaida, K. Watanabe, and K. D. A. Wendt. Determination of ionization potential of calcium by high-resolution resonance ionization spectroscopy. *Journal of the Physical Society of Japan*, 75(3):034302–1–034302–10, 2006. doi: 10.1143/JPSJ.75.034302.

P. J. Mohr, B. N. Taylor, and D. B. Newell. Codata recommended values of the fundamental physical constants: 2010. *Reviews of Modern Physics*, 84(4):1527–1605, October-December 2012.

M. Morgan, S. Anders, M. Lawrence, P. Aboyoun, H. Pagès, and R. Gentleman. ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, 25:2607–2608, 2009. doi: 10.1093/bioinformatics/btp450. URL `http://dx.doi.org10.1093/bioinformatics/btp450`.

M. G. Morgan and M. Henrion. *Uncertainty — A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, New York, NY, first paperback edition, 1992. 10th printing, 2007.

D. E. Morris, J. E. Oakley, and J. A. Crowe. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4, 2014.

P. K. Moser. Belief. In R. Audi, editor, *The Cambridge Dictionary of Philosophy*, pages 78–79. Cambridge University Press, Cambridge, UK, second edition, 1999. 11th printing, 2009.

F. Mosteller and J. W. Tukey. *Data Analysis and Regression*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1977.

NASA. *Estimation and Evaluation of Measurement Decision Risk*. National Aeronautics and Space Administration, Washington DC 20546, July 2010. NASA Measurement Quality Assurance Handbook (Annex 4), NASA-HDBK-8739.19-4.

J. A. Nelder and R. Mead. A simplex algorithm for function minimization. *Computer Journal*, 7:308–313, 1965.

J. V. Nicholas and D. R. White. *Traceable Temperatures*. John Wiley & Sons, Chichester, England, second edition, 2001.

D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multi-resolution gaussian process model for the analysis of large spatial data sets. NCAR Technical Note NCAR/TN-504+STR, National Center for Atmospheric Research, Boulder, Colorado, December 2013.

D. Nychka, D. Hammerling, S. Sain, and N. Lenssen. *LatticeKrig: Multiresolution Kriging based on Markov random fields*, 2014. URL `http://CRAN.R-project.org/package=LatticeKrig`. R package version 3.4.

A. O'Hagan. SHELF: the Sheffield Elicitation Framework, 2012. URL `www.tonyohagan.co.uk/shelf`. Version 2.0.

A. O'Hagan. Eliciting and using expert knowledge in metrology. *Metrologia*, 51(4):S237–S244, 2014. doi: 10.1088/0026-1394/51/4/S237.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. Statistics in Practice. John Wiley & Sons, Chichester, England, 2006. ISBN: 978-0-470-02999-2.

OLES. *Ballistic Resistance of Body Armor, NIJ Standard-0101.06*. National Institute of Justice, Washington, DC, July 2008. Office of Law Enforcement Standards, National Institute of Standards and Technology.

K. A. Olive and Particle Data Group. Review of particle physics. *Chinese Physics C*, 38(9):090001, 2014. doi: 10.1088/1674-1137/38/9/090001.

C. Osborne. Statistical calibration: A review. *International Statistical Review*, 59(3):309–336, December 1991.

E. Pebesma, D. Cornford, G. Dubois, G.B.M. Heuvelink, D. Hristopoulos, J. Pilz, U. Stoehlker, G. Morin, and J.O. Skoien. INTAMAP: the design and implementation of an interoperable automated interpolation web service. *Computers & Geosciences*, 37:343–352, March 2010. doi: 10.1016/j.cageo.2010.03.019.

H. S. Peiser, N. E. Holden, P. De Bièvre, I. L. Barnes, R. Hagemann, J. R. de Laeter, T. J. Murphy, E. Roth, M. Shima, and H. G. Thode. Element by element review of their atomic weights. *Pure and Applied Chemistry*, 56(6):695–768, 1984. URL `http://dx.doi.org/10.1351/pac198456060695`. International Union Of Pure and Applied Chemistry, Inorganic Chemistry Division, Commission On Atomic Weights and Isotopic Abundances.

L. A. B. Pilkington. Review lecture: The float glass process. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 314(1516):1–25, 1969. doi: 10.1098/rspa.1969.0212.

J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects*

Models, 2014. URL `http://CRAN.R-project.org/package=nlme`. R package version 3.1-115.

J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-Plus*. Springer-Verlag, New York, NY, 2000.

J. C. Pinheiro, C. Liu, and Y. N. Wu. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t* distribution. *Journal of Computational and Graphical Statistics*, 10(2):249–276, 2001.

S. J. Pocock and D. J. Spiegelhalter. Domiciliary thrombolysis by general practitioners. *British Medical Journal*, 305(6860):1015–1015, 1992. doi: 10.1136/bmj.305.6860.1015.

A. Possolo. Copulas for uncertainty analysis. *Metrologia*, 47:262–271, 2010.

A. Possolo. Model-based interpolation, prediction, and approximation. In A. M. Dienstfrey and R. F. Boisvert, editors, *Uncertainty Quantification in Scientific Computing*, IFIP Advances in Information and Communications Technology, pages 195–211. Springer, New York, 2012. 10th IFIP WG 2.5 Working Conference, WoCoUQ 2011, Boulder, CO, USA, August 1–4, 2011.

A. Possolo and C. Elster. Evaluating the uncertainty of input quantities in measurement models. *Metrologia*, 51(3):339–353, June 2014. doi: 10.1088/0026-1394/51/3/339.

A. Possolo and 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, November 2011. doi: 10.13140/RG.2.1.2256.8482. URL `www.itl.nist.gov/div898/possolo/TutorialWEBServer/TutorialMetrologists2011Nov09.xht`.

J. B. Quinn and G. D. Quinn. A practical and systematic review of Weibull statistics for reporting strengths of dental materials. *Dental Materials*, 26:135–147, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York, NY, second edition, 2007.

P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, and M. Maechler. *robustbase: Basic Robust Statistics*, 2012. URL `http://CRAN.R-project.org/package=robustbase`. R package version 0.9-7.

T. Rowan. *Functional Stability Analysis of Numerical Algorithms*. PhD thesis, University of Texas at Austin, Austin, TX, 1990. Department of Computer Sciences.

A. L. Rukhin. Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):451–469, 2013. doi: 10.1111/j.1467-9868.2012.01047.x.

A. L. Rukhin and A. Possolo. Laplace random effects models for interlaboratory studies. *Computational Statistics and Data Analysis*, 55:1815–1827, 2011.

A. L. Rukhin and M. G. Vangel. Estimation of a common mean and weighted means statistics. *Journal of the American Statistical Association*, 93:303–308, 1998.

D. Rumble, S. Bowring, T. Iizuka, T. Komiya, A. Lepland, M. T. Rosing, and Y. Ueno. The oxygen isotope composition of earth's oldest rocks and evidence of a terrestrial magma ocean. *Geochemistry, Geophysics, Geosystems*, 14(6):1929–1939, 2013. doi: 10.1002/ggge.20128.

M. Schantz and S. Wise. CCQM–K25: Determination of pcb congeners in sediment. *Metrologia*, 41(*Technical Supplement*):08001, 2004.

E. Schwitzgebel. Belief. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information (CSLI), Stanford University, summer edition, 2015.

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

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52 (3,4):591–611, 1965.

C. P. Sison and J. Glaz. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, 90(429):366–369, 1995. doi: 10.1080/01621459.1995.10476521.

S.M. Smith. *National Geochemical Database: Reformatted data from the National Uranium Resource Evaluation (NURE) Hydrogeochemical and Stream Sediment Reconnaissance (HSSR) Program*, 2001. URL `http://greenwood.cr.usgs.gov/pub/open-file-reports/ofr-97-0492/index.html`. Version 1.30.

D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Statistics in Practice. John Wiley Sons, Chichester, England, 2004.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Verlag, New York, NY, 1999.

M. Steup. Epistemology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, California, spring 2014 edition, 2014. URL `http://plato.stanford.edu/archives/spr2014/entries/epistemology/`.

J. A. Stone and J. H. Zimmerman. Refractive index of air calculator. In T. Doiron, editor, *Engineering Metrology Toolbox*. National Institute of Standards and Technology, Gaithersburg, MD, 2011. URL `http://emtoolbox.nist.gov/`.

Student. On the error of counting with a haemacytometer. *Biometrika*, 5(3):351–360, February 1907.

S. S.-C. Tai, M. Bedner, and K. W. Phinney. Development of a candidate reference measurement procedure for the determination of 25-hydroxyvitamin D3 and 25-hydroxyvitamin D2 in human serum using isotope-dilution liquid chromatographyâĹŠtandem mass spectrometry. *Analytical Chemistry*, 82(5):1942–1948, 2010. doi: 10.1021/ac9026862.

B. N. Taylor and C. E. Kuyatt. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. National Institute of Standards and Technology, Gaithersburg, MD, 1994. URL `http://physics.nist.gov/Pubs/guidelines/TN1297/tn1297s.pdf`. NIST Technical Note 1297.

The ALEPH Collaboration, The DELPHI collaboration, the L3 collaboration, the OPAL Collaboration, and The LEP Electroweak Working Group. Electroweak measurements in electron-positron collisions at w-boson-pair energies at lep. *Physics Reports*, 532(4):119–244, November 2013. doi: 10.1016/j.physrep.2013.07.004.

M. Thompson. What *exactly* is uncertainty? *Accreditation and Quality Assurance*, 17:93–94, 2011.

M. Thompson and S. L. R. Ellison. Dark uncertainty. *Accreditation and Quality Assurance*, 16:483–487, 2011.

B. Toman. Bayesian approaches to calculating a reference value in key comparison experiments. *Technometrics*, 49(1):81–87, February 2007.

B. Toman and A. Possolo. Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance*, 14:553–563, 2009.

B. Toman and A. Possolo. Erratum to: Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance*, 15:653–654, 2010.

C. M. Tsui, A. Y. K. Yang, and H. W. Li. Software tools for evaluation of measurement models for complex-valued quantities in accordance with Supplement 2 to the GUM. *NCSLI Measure Journal of Measurement Science*, 7(3):48–55, September 2012.

M.P.J. van der Loo. The stringdist package for approximate string matching. *The R Journal*, 6, 2014. URL `http://CRAN.R-project.org/package=stringdist`. Accepted for publication.

W. Viechtbauer. Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26:37–52, 2007.

W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36 (3):1–48, 2010. URL `http://www.jstatsoft.org/v36/i03/`.

P. J. Villacorta. *MultinomialCI: Simultaneous confidence intervals for multinomial proportions according to the method by Sison and Glaz*, 2012. URL `\url{http://CRAN.R-project.org/package=MultinomialCI}`. R package version 1.0.

J. Wallace. Ten methods for calculating the uncertainty of measurement. *Science & Justice*, 50(4):182–186, 2010. ISSN 1355-0306. doi: 10.1016/j.scijus.2010.06.003.

L. Wasserman. *All of Statistics, A Concise Course in Statistical Inference*. Springer Science+Business Media, New York, NY, 2004.

R. E. Weston Jr. Anomalous or mass-independent isotope effects. *Chemical Reviews*, 99(8):2115–2136, 1999. doi: 10.1021/cr9800154.

J. R. Whetstone, W. G. Cleveland, G. P. Baumgarten, S. Woo, and M. C. Croarkin. Measurements of coefficients of discharge for concentric flange-tapped square-edged orifice meters in water over the reynolds number range 600 to 2,700,000. NIST Technical Note 1264, National Institute of Standards and Technology, Gaithersburg, MD, June 1989. U.S. Department of Commerce.

G. H. White and I. Farrance. Uncertainty of measurement in quantitative medical testing — a laboratory implementation guide. *Clinical Biochemistry Reviews*, 25 Supplement (ii):S1–S24, November 2004.

R. White. The meaning of measurement in metrology. *Accreditation and Quality Assurance*, 16:31–41, 2011. doi: 10.1007/s00769-010-0698-1.

M. E. Wieser, N. Holden, T. B. Coplen, J. K. Böhlke, M. Berglund, W. A. Brand, P. De Bièvre, M. Gröning, R. D. Loss, J. Meija, T. Hirata, T. Prohaska, R. Schoenberg, G. O'Connor, T. Walczyk, S. Yoneda, and X.-K. Zhu. Atomic weights of the elements 2011 (IUPAC Technical Report). *Pure and Applied Chemistry*, 85(5): 1047–1078, 2013. URL `http://dx.doi.org/10.1351/PAC-REP-13-03-02`.

Wikipedia. Probability distribution — Wikipedia, The Free Encyclopedia, 2015. URL `http://en.`

wikipedia.org/w/index.php?title=Probability_distribution&oldid=647572820. [Online; accessed 14-April-2015].

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.

M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, March 1968.

S. A. Wise and R. L. Watters. *Standard Reference Materials 8535 (VSMOW), 8536 (GISP), 8537 (SLAP)*. Office of Reference Materials, National Institute of Standards and Technology, Department of Commerce, Gaithersburg, Maryland, 2005a. URL http://www.nist.gov/srm/. In cooperation with the International Atomic Energy Agency.

S. A. Wise and R. L. Watters. *Standard Reference Material 1d, Argillaceous Limestone*. Office of Reference Materials, National Institute of Standards and Technology, Department of Commerce, Gaithersburg, Maryland, 2005b. URL http://www.nist.gov/srm/.

S. N. Wood. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95–114, 2003.

S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL, 2006.

World Meteorological Organization. *Guide to Meteorological Instruments and Methods of Observation*. World Meteorological Organization, Geneva, Switzerland, seventh edition, 2008. WMO-No. 8.

E. D. Young, A. Galy, and H. Nagahara. Kinetic and equilibrium mass-dependent isotope fractionation laws in nature and their geochemical and cosmochemical significance. *Geochimica et Cosmochimica Acta*, 66(6): 1095–1104, 2002.

J. Ypma. Introduction to `nloptr`: an R interface to NLopt, August 2014. URL http://cran.fhcrc.org/web/packages/nloptr/. Vignette for R package `nloptr`.