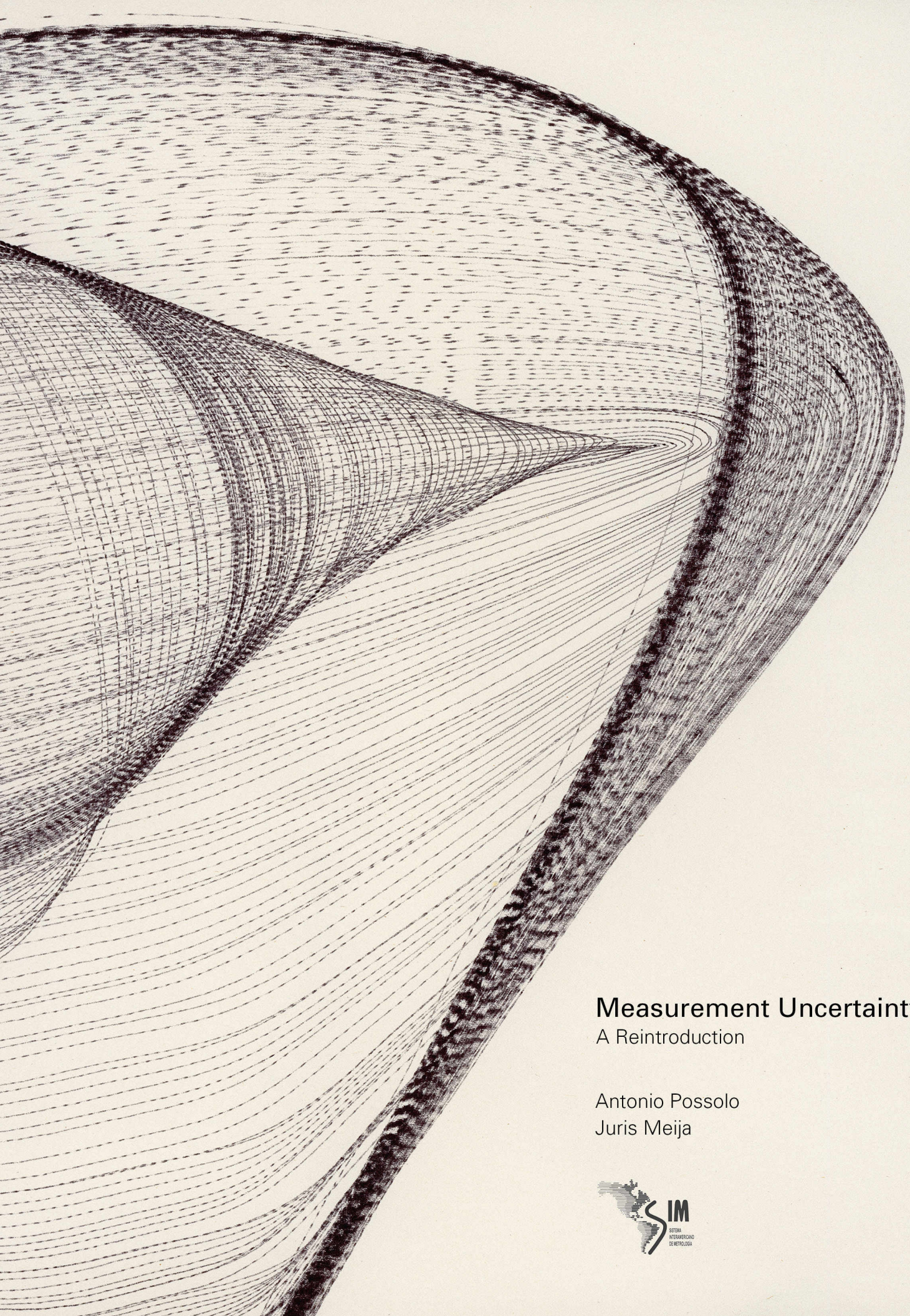


Measurement Uncertainty:
A Reintroduction

Antonio Possolo
Juris Meija

SIM
SISTEMA INTERAMERICANO DE METROLOGIA

Antonio Possolo & Juris Meija

Measurement Uncertainty: A Reintroduction

sistema interamericano de metrologia — sim

Published by SISTEMA INTERAMERICANO DE METROLOGIA — SIM
Montevideo, Uruguay: September 2020

ISBN 978-0-660-36124-6
DOI 10.4224/40001835

Cover design by Neeko Paluzzi, including original, personal artwork © 2016 Juris Meija. Front matter,
body, and end matter composed with LaTeX as implemented in the 2020 TeX Live distribution from the
TeX Users Group, using the `tufte-book` class (© 2007-2015 Bil Kleb, Bill Wood, and Kevin Godby), with
Palatino, Helvetica, and Bera Mono typefaces.

Knowledge would be fatal. It is the uncertainty that charms one. A mist makes things wonderful.

— Oscar Wilde (1890, *The Picture of Dorian Gray*)

Preface

Our original aim was to write an introduction to the evaluation and expression of measurement uncertainty as accessible and succinct as Stephanie Bell's little jewel of *A Beginner's Guide to Uncertainty of Measurement* [Bell, 1999], only showing a greater variety of examples to illustrate how measurement science has grown and widened in scope in the course of the intervening twenty years.

The recent, very welcome *Introduction to Measurement Uncertainty* that Blair Hall and Rod White have made available to the community [Hall and White, 2018], occupies a middle ground in terms of complexity. It presents two realistic examples in considerable detail (using a ruler, and calibrating a thermometer), and it excels in typographical design, from which we have obviously drawn inspiration.

We have organized our narrative so that a reader who is primarily interested in weighing may skip the discussion of counting, and similarly for all the other sections. Even subsections within the same section can, in most cases, be read independently of one another: for example, to learn how to compare two measurement methods, while remaining unconcerned with how to compare a measured value with a corresponding certified value.

While some of our examples are very simple and likely to appeal to a broad audience (measuring the volume of a storage tank, or surveying a plot of land), others may interest only a more narrowly specialized sector of the readership (measuring abortion rates, or calibrating a resistor using a Wheatstone bridge). Some may appear, at first blush, to be narrowly focused (measuring the Hubble-Lemaître constant), but in fact employ techniques that are widely applicable. Still others are fairly complex, yet are likely to draw the attention of many readers (calibrating a GC-MS system, or averaging models for a flu epidemic).

The predominant approach to measurement uncertainty involves probabilistic concepts and requires the application of statistical methods. We have chosen not to hide the attending difficulties, and have striven instead to explain the models we use, and the calculations necessary to apply them, in fair detail, providing computer codes to carry them out.

These technicalities, no matter how clearly one may be able to explain them, inevitably will be challenging obstacles for many readers. Two appendices, one on probability, the other on statistics, may help motivated readers familiarize themselves with concepts and methods sufficiently to overcome such obstacles, yet they demand considerable commitment from the reader.

We have illustrated the application of a wide range of statistical models and methods, some from the classical school, others of a Bayesian flavor, especially when it is advantageous to incorporate preexisting knowledge about a measurand. However, the choice of school or flavor is not critical.

The key resolution is to approach each problem with flexibility, being deferential to the data and attentive to the purpose of the inquiry: to select models and employ data reduction techniques that are verifiably adequate for the data in hand; to give each problem a custom solution tailored for the purpose that the results are intended to serve; all along heeding Lincoln Moses's advice that "You have to have a good data-side manner."

Acknowledgments & Disclaimers

The authors are honored by the support the SIM Council has chosen to grant to this contribution by approving its publication under the SIM banner: Claire Saundry (NIST), *President*; Hector Laiz (INTI), *Former President*; J. Salvador Echeverria Villagomez (CENAM), *Technical Committee Chair*; Sally Bruce (NIST), QSTF *Chair*; Javier Arias (CENAMEP AIP), *Project Coordinator*; Rodrigo Costa-Felix (INMETRO), *Professional Development Coordinator*; Edwin Arvey Cristancho-Pinilla (INM), ANDIMET *Coordinator*; Pedro Ibarra (INN), SURAMET *Coordinator*; Claudia Estrada (CIM), CAMET *Coordinator*; I-Ronn Audain (SKNBS), CARIMET *Coordinator*; and Victor Lizardi (CENAM), NORAMET *Coordinator*.

The example involving a Wheatstone bridge is patterned after a laboratory experiment of Physics 321 as taught by Dan McCammon and Vincent Liu at the University of Wisconsin-Madison, in the fall of 2016.

John Sieber, Tom Vetter, and Adam Pintar, all from NIST shared the observations of fluorescence intensity used to illustrate a permutation test for homogeneity of a reference material. George D. Quinn (Material Measurement Laboratory, NIST), kindly provided a set of determinations of rupture strength of alumina coupons. Laura Wood and Katherine Rimmer (Material Measurement Laboratory, NIST) gave us permission to use determinations of the mass fraction of arsenic in kudzu.

Several colleagues from SIM countries provided very valuable comments and suggestions for improvement, and detected errors, in an early draft. We are particularly grateful to Hugo Gasca Aragón (CENAM, Mexico), Bryan Calderón (LACOMET, Costa Rica), Silvina Forastieri (INTI, Argentina), Hari Iyer (NIST), Dianne Lalla-Rodrigues (ABBS, Antigua and Barbuda), Wilson Naula (INEN, Ecuador), Claudia Santo (SIM), and Barry Wood (NRC Canada), for their generous contributions and perceptive insights.

The *Statistics* Appendix suggests that statistics is an art that one learns from master artisans. Antonio Possolo was fortunate to have apprenticed with John Hartigan (Yale), Frank Anscombe (Yale), and John Tukey (Princeton).

Navigation

Every word or number that is of a color other than black, except any so colored that appears in figures, is a clickable hyperlink: wild strawberry hyperlinks lead to related material in the main body of the document or in the appendices; process blue hyperlinks point to bibliographic references; and forest green hyperlinks take the reader to a web page external to this document.

Mentions of NIST reference materials are implied references to their certificates, which can be obtained by entering the corresponding reference number in the field labeled "SRM/RM Number" in the web page for NIST STANDARD REFERENCE MATERIALS.

Contents

Measurement

Our ancestors were shepherds that counted sheep, surveyors that sized agricultural land, traders that weighed gold pieces, timekeepers that relied on sundials, merchants that graded silk according to its fineness, and healers that assigned medicinal plants to categories reflecting their powers (*cf.* Todd [1990]).

Counting, surveying, weighing, timing, ranking, and classifying all serve to assign a value to a property (*measurand*) of an object of interest, and all are instances of measurement provided they satisfy these requirements: (i) the assignment of value is based on comparison with a standard that is recognized as a common reference by the community of producers and users of the measurement result; (ii) the measured value is qualified with an evaluation of measurement uncertainty whose practical meaning is well understood and agreed upon; (iii) the measurement result (measured value together with its associated measurement uncertainty) is used to inform an action or decision.

A measured value is an estimate of the true value of a property, which may be quantitative or qualitative. Counting, surveying, weighing, and timing all produce estimates of quantitative measurands. Ranking applies to qualities that come by degrees that can meaningfully be ordered from smallest to largest, or weakest to strongest (for example, the Mohs hardness of a mineral, or the spiciness of a curry). Classification (or identification) assigns objects to categories that are either identical or different, but that cannot otherwise be ordered or quantified (for example, the identity of the nucleobase at a particular location of a DNA strand, or the gender of an athlete).

In ancient Egypt, measurement was considered important even in the afterlife: Anubis (god of death) leads the scribe Hunefer to judgement, where his heart is weighed against the Feather of Truth. Thoth (god of writing) records the result, while Ammit, Devourer of the Dead, awaits the verdict.
— *Book of the Dead* (1275 BCE) British Museum (EA 9901,3)

Recognizing and quantifying the uncertainty that invariably clouds our knowledge of the world is a hallmark of science. It informs actions and decisions in all fields of the human endeavor: protecting against incoming storms, planning crops, responding to epidemics, or managing industrial inventories. Measurement uncertainty is an integral part of every measurement result, characterizing its quality.

Measurement Uncertainty

Measurement uncertainty is the doubt about the true value of the measurand that remains after making a measurement [Possolo, 2015]. The corresponding margin of doubt is characterized by its width (size of the uncertainty) and by its depth (severity of the uncertainty): the wider this margin, the larger the range of values of the measurand that are consistent with the measured value; the deeper this margin, the smaller the confidence that the true value of the measurand indeed lies within that margin [Bell, 1999].

There is no science without measurements, no quality without testing, and no global commerce without standards. Since no measurement is perfect, evaluating measurement uncertainty and taking it into account are prerequisites for interpreting and using measurement results.

Uncertainty often originates not only from imperfections in measurement, but also from the natural variability of the true values of the properties we seek to measure. For example, the exact amount of aspirin may vary slightly among nominally identical pills, and the volume of dishwashing liquid in nominally identical bottles often varies enough to be perceptible to the naked eye.

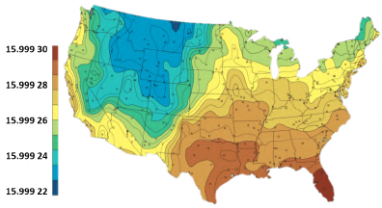
In addition to imperfect measurements or natural variability of the true values of measurands, it is fairly common for there to be ambiguity, or incomplete specification, of the very definition of what we are trying to measure. The following three examples describe cases where such ambiguity was an important source of uncertainty.

In January, 2015, the U.S. Supreme Court[1] decided a case where the basic dispute concerned the meaning of the term "molecular weight" as it had been used in a patent filed by Teva. The Court considered that "the term might refer to the weight of the most numerous molecule, it might refer to weight as calculated by the average weight of all molecules, or it might refer to weight as calculated by an average in which heavier molecules count for more."

Driving under the influence (DUI) court cases rely on measurements made to determine whether alcohol concentration exceeded 0.08 g per 100 mL of blood, or 0.08 g per 210 L of breath. Typically, the prosecution has to demonstrate that the alcohol concentration indeed exceeded the 0.08 level beyond reasonable doubt, which is often taken to mean 99 % confidence.

Besides the sizable measurement uncertainty, which in large part is attributable to calibration uncertainty,[2] the factors affecting the outcome of breath tests include body temperature, blood makeup (hematocrit, the volume fraction of red blood cells in the blood), or the manner of breathing. Moreover, uncertainty can surround

Truth lies hidden in a castle's keep, surrounded by uncertainty, which is represented by the moat. The width of the moat portrays the margin of doubt, and its depth illustrates the severity of the doubt [Bell, 1999] (Almourol Castle, Portugal — Wikimedia Commons, Daniel Feliciano, 2003).



The speed of light in vacuum has exactly one true value that is invariant in time and space, according to the prevailing view of the universe. But the true value of the atomic weight of oxygen varies significantly across USA river waters, reflecting the spatial variability of the amount fractions of its different isotopes [Kendall and Coplen, 2001].

[1] *Teva Pharmaceuticals USA, Inc.* v. *Sandoz, Inc.* 574 U. S. 318 (2015), 2015

) Case No. C076949 and 9Y6231062
)
) ORDER SUPPRESSING DEFENDANT'S
) BREATH-ALCOHOL MEASUREMENTS IN
) THE ABSENCE OF A MEASUREMENT
) FOR UNCERTAINTY

Measurement uncertainty is crucial to determining whether laws are broken (excerpt from a 2010 King County District Court ruling, Washington, USA).

[2] S. Cowley and J. Silver-Greenberg. These Machines Can Put You in Jail. Don't Trust Them. *The New York Times*, November 3, 2019. Business Section; and J. Silver-Greenberg and S. Cowley. 5 Reasons to Question Breath Tests. *The New York Times*, November 3, 2019. Business Section

many other aspects of the measurement: some parts of the body will have higher blood-alcohol concentration than others, with the alcohol levels in arterial and venous blood possibly differing by as much as a factor of two [Simpson, 1987]. Even the very definition of alcohol, surprisingly can include not only ethanol but also other low molecular weight alcohols such as methanol or isopropanol.

Defining gender, in particular of athletes participating in sports where men and women compete separately, has become a vivid instance of definitional uncertainty, as the understanding has widened, among biologists, that the binary notion of gender (male or female) is an inaccurate oversimplification. In fact, gender is a spectrum,[3] for there are several different ways in which its value may be expressed or assigned — based on anatomical features, hormonal profile, chromosomal structure, or self-identification —, which may contradict each other, giving rise to uncertainty.

The foregoing served to highlight how consideration of measurement uncertainty pervades not only areas of science and technology, but also many aspects of everyday life. Next we illustrate how measurement uncertainty associated with measured values can be propagated to the results of simple calculations involving these values.

[3] C. Ainsworth. Sex redefined. *Nature*, 518:288–291, February 2015. doi:10.1038/518288a. News Feature

Sums, Products, and Ratios

In many cases, quantities of interest are expressed as sums, products, or ratios of other quantities that may have been measured previously or that are measured in the course of the measurement experiment. Such fairly simple measurement models serve to illustrate the basic procedures involved in uncertainty evaluations, including the propagation of uncertainties from input quantities to an output quantity, as in the following three examples: (i) the plasma anion gap (expressed as a sum of four measured amount concentrations); (ii) the volume of a cylindrical storage tank (expressed as a product of two measured lengths); and (iii) the resistance of an electric resistor (which is given by a ratio involving several measured resistances). In this third example we will also illustrate the use of the *NIST Uncertainty Machine*.[4]

[4] T. Lafarge and A. Possolo. The NIST Uncertainty Machine. *NCSLI Measure Journal of Measurement Science*, 10(3):20–27, September 2015. doi:10.1080/19315775.2015.11721732

Plasma Anion Gap

The plasma anion gap, Δc_{AG}, is used in clinical biochemistry to determine whether there is an imbalance of electrolytes in the blood, which may be a result of diabetes or of kidney disease, among other possiblities. It is defined as a linear combination of the amount concentration of two cations and two anions:

$$\Delta c_{\mathrm{AG}} = c(\mathrm{Na}^+) + c(\mathrm{K}^+) - c(\mathrm{Cl}^-) - c(\mathrm{HCO}_3^-).$$

There are several different definitions of the anion gap. For example, it is common to omit potassium or to include corrections due to albumin.

Consider the values that were measured for a particular patient, shown in the table alongside. For this patient,

$$\Delta c_{\mathrm{AG}} = (137 + 4 - 106 - 10)\,\mathrm{mmol/L} = 25\,\mathrm{mmol/L},$$

which generally would be regarded as being of clinical concern. However, the interpretation of any result of laboratory medicine requires consideration of the complete clinical profile of the patient,[5] and requires also that measurement uncertainty be taken into account.

[5] G. H. White, C. A. Campbell, and A. R Horvath. Is this a Critical, Panic, Alarm, Urgent, or Markedly Abnormal Result? *Clinical Chemistry*, 60(12):1569–1570, December 2014. doi:10.1373/clinchem.2014.227645

The uncertainty associated with the value of Δc_{AG} is determined by the reported uncertainties for the individual ion amount concentrations. These are the sizes of the margins of uncertainty discussed above, under *Measurement Uncertainty*. White [2008] does not describe how they were evaluated, or which sources of uncertainty may have contributed to these values, but refers to them as standard deviations.

This suggests that the underlying model for the measured amount concentrations involves random variables and probability distributions, which provides a way forward to evaluate the standard uncertainty of the anion gap.

ion	c	$u(c)$
Na^+	137	1.48
K^+	4	0.04
Cl^-	106	0.72
HCO_3^-	10	0.84

Amount concentrations of ions (mmol/L) that were measured for a particular patient [White, 2008].

Indeed, if those four amount concentrations can be regarded as outcomes of independent random variables, then Δc_{AG} also is a random variable because it is a function of these random variables. Its variance, denoted $u^2(\Delta c_{\mathrm{AG}})$ below, can be computed exactly because the AG is a linear combination of the four amount concentrations, and the corresponding standard deviation, which will become its standard uncertainty, $u(\Delta c_{\mathrm{AG}})$, is the square root of this variance:

If an output quantity $Y = \alpha_1 X_1 + \cdots + \alpha_n X_n$ is a linear combination of uncorrelated input quantities for which estimates $x_1, \ldots, x_n$ and associated standard uncertainties $u(x_1), \ldots, u(x_n)$ are available, $\alpha_1, \ldots, \alpha_n$ are known constants, and $y = \alpha_1 x_1 + \cdots + \alpha_n x_n$, then $u^2(y) = \alpha_1^2 u^2(x_1) + \cdots + \alpha_n^2 u^2(x_n)$

$$\begin{aligned} u^2(\Delta c_{\mathrm{AG}}) &= u^2(c(\mathrm{Na}^+)) + u^2(c(\mathrm{K}^+)) + u^2(c(\mathrm{Cl}^-)) + u^2(c(\mathrm{HCO}_3^-)) \\ &= (1.48\,\mathrm{mmol/L})^2 + (0.04\,\mathrm{mmol/L})^2 + \\ &\quad (0.72\,\mathrm{mmol/L})^2 + (0.84\,\mathrm{mmol/L})^2 \\ &= (1.85\,\mathrm{mmol/L})^2 \end{aligned}$$

Even though Δc_{AG} involves sums and differences, the variances of the quantities being added or subtracted are all added.

The precise meaning of $u(\Delta c_{\mathrm{AG}}) = 1.85\,\mathrm{mmol/L}$ depends on the probability distribution of the random variable that is being used as a model for Δc_{AG}. If the four ion concentrations were modeled as Gaussian (or, normal) random variables, then so would be the Δc_{AG}, because a linear combination of independent Gaussian random variables is also Gaussian. In these circumstances, the conclusion would be that the true value of the Δc_{AG} is 25 mmol/L to within 1.85 mmol/L, with approximately 68 % probability.

It is a surprising fact that, for many distributions that the Δc_{AG} may have, the interval $\Delta c_{\mathrm{AG}} \pm 2u(\Delta c_{\mathrm{AG}})$ will include the true value of Δc_{AG} with approximately 95 % probability [Freedman et al., 2007].

Volume of Storage Tank

Consider the problem of evaluating and expressing the uncertainty that surrounds the internal volume V of a cylindrical storage tank, derived from measurement results for its radius R, and for its height H. Since the volume is a nonlinear function of the radius and height, $V = \pi R^2 H$, the form of calculation illustrated for the anion gap does not apply to this case.

The radius was measured by climbing a set of stairs to the tank's roof, whose shape and size are essentially identical to its base, measuring its diameter with a tape, and reporting the estimate of the radius as 8.40 m, give or take 0.03 m. This "give or take" is the margin of uncertainty, but without additional information it is not particularly meaningful or useful: one needs to know, for example, how likely the true value is of lying between 8.37 m and 8.43 m. That is, one needs to be able to translate uncertainty into probability. This is often done by regarding the measured value, 8.40 m, as the observed value of a random variable whose mean is the true value of R, and whose standard deviation is 0.03 m. This interpretation motivates calling the "give or take" *standard uncertainty*.

The volume of a cylindrical, oil storage tank is a non-linear function of its height and diameter — PixelSquid (use licensed 2020).

In order to define the meaning of "give or take" fully we need to specify what the 0.03 m actually subsumes, that is, which sources of uncertainty contribute to it, how they may have been evaluated, and how their contributions may have been combined. In addition, we must specify how likely it is that the true value of the radius indeed lies within 0.03 m of the measured value, 8.40 m.

What does the 0.03 m *actually subsume?* The standard uncertainty should reflect contributions from all recognized sources of uncertainty.

- Some of these contributions originate in the tape itself (how and when it was calibrated, or the effect of temperature on the tape);
- Other contributions derive from how the tape will have been laid out along a diameter of the roof (how stretched it may have been, how closely it will have passed to the actual center of the roof, and whether it touched and went over any rivets or ridges that may have made it deviate from a straight line parallel to the roof);
- Still other effects are attributable to how the tape was used by the person making the measurement (whether multiple measurements were made of the length of the diameter, and if so whether they were averaged or combined in some other way);
- And there will also be contributions from sources that are specific to the tank itself (how close to a perfect circle its roof may be, or how the temperature may affect the tank's size and shape)

How likely is it that the true value of the radius indeed lies within 0.03 m *of the measured value,* 8.40 m*?* To answer this question one needs a particular model for the uncertainty that the question alludes to. The kind of model used most often in metrology to address this question is a probabilistic model that characterizes in sufficient detail the random variable mentioned above. Such model is a *probability distribution*.

But which probability distribution? The answer depends on what is known about the sources of uncertainty listed above, and on how their contributions will have been combined into the reported margin of uncertainty. A common choice (but by no means the best in all cases) is to use a Gaussian distribution as the model that lends meaning to the margin of uncertainty. In such case one can claim that the probability is about 68 % that the true value of the radius is within 0.03 m of its measured value.

The same questions need to be answered, and comparable modeling assumptions need to be made for the tank's height, H, which was measured using a plumb line dropped from the edge of the roof to

the concrete platform that the tank is anchored to. The result turned out to be 32.50 m give or take 0.07 m. The estimate of the volume is $V = \pi R^2 H = 7204\,\mathrm{m}^3$.

The measurement model, $V = \pi R^2 H$, expresses the output quantity V as a function of the two input quantities, R and H, whose values are surrounded by uncertainty. If, for the purposes of uncertainty evaluation, both R and H are modeled as random variables, then V will also be a random variable and the problem of evaluating its uncertainty can be solved either by characterizing its probability distribution fully, or, at a minimum, by computing its standard deviation.

We'll do both under the assumption that R and H are independent random variables, and that both have Gaussian distributions centered at their measured values, with standard deviations equal to their standard uncertainties.

Gauss's formula[6] [Possolo and Iyer, 2017, VII.A.2], which is used in the *Guide to the expression of uncertainty in measurement* (GUM) [JCGM 100:2008], provides a practicable alternative that will produce a particularly simple approximation to the standard deviation of the output quantity because it is a product of powers of the input quantities: $V = \pi R^2 H^1$. The approximation is this

$$\left(\frac{u(V)}{V}\right)^2 \approx \left(2 \times \frac{u(R)}{R}\right)^2 + \left(1 \times \frac{u(H)}{H}\right)^2.$$

Note that π does not figure in this formula because it has no uncertainty, and that the "2" and the "1" that appear as multipliers on the right-hand side are the exponents of R and H in the formula for the volume. The approximation is likely to be good when the relative uncertainties, $u(R)/R$ and $u(H)/H$, are small — say, less than 10 % —, as they are in this case. Therefore

$$u(V) \approx 7204\,\mathrm{m}^3 \sqrt{\left(2 \times \frac{0.03\,\mathrm{m}}{8.40\,\mathrm{m}}\right)^2 + \left(\frac{0.07\,\mathrm{m}}{32.50\,\mathrm{m}}\right)^2} = 54\,\mathrm{m}^3.$$

A Monte Carlo method [Possolo and Iyer, 2017, VII.A.3] for uncertainty propagation introduced by Morgan and Henrion [1992] and described in JCGM 101:2008, provides yet another eminently practicable alternative, whose validity does not depend on the relative standard uncertainties being small. The idea and execution both are very simple, like this:

(1) Make a large number ($K \approx 10^6$) of drawings from the probability distributions of R and H, using their measured values as

[6] C. Gauss. Theoria combinationis observationum erroribus minimis obnoxiae. In *Werke, Band IV, Wahrscheinlichkeitsrechnung und Geometrie*. Königlichen Gesellschaft der Wissenschaften, Göttingen, 1823. URL http://gdz.sub.uni-goettingen.de

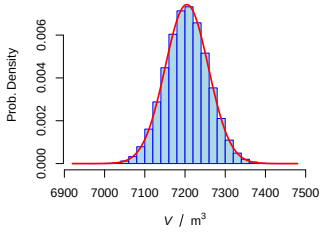
In general, if the measurement model expresses the output quantity as $Y = \kappa X_1^{\alpha_1} \ldots X_n^{\alpha_n}$, with mean η and standard deviation τ, and $X_1, \ldots, X_n$ are independent random variables with means $\mu_1, \ldots, \mu_n$ and standard deviations $\sigma_1, \ldots, \sigma_n$, such that $\sigma_1/\mu_1, \ldots, \sigma_n/\mu_n$ are small (say, $< 10\,\%$), and $\alpha_1, \ldots, \alpha_n$ are constants, then $(\tau/\eta)^2 \approx (\alpha_1\sigma_1/\mu_1)^2 + \cdots + (\alpha_n\sigma_n/\mu_n)^2$.

the means of these distributions, and their reported standard uncertainties as the standard deviations.

(2) For each pair of these draws, r_k and h_k, calculate the volume of the cylinder $v_k = \pi r_k^2 h_k$, for $k = 1, \ldots, K$.

(3) Calculate the average of these values of the volume, $v_1, \ldots, v_K$, and use it as an estimate of the mean value of V, and their standard deviation as an estimate of $u(V)$.

Using samples of size $K = 10^6$, we reached the conclusion that $V = 7204\,\mathrm{m}^3$, give or take $54\,\mathrm{m}^3$, and the histogram of these one million replicates shows that V has a probability density that is virtually indistinguishable from the density of a Gaussian distribution with this mean and standard deviation. Note, however, that in general the probability distribution of the output quantity need not be close to Gaussian, even when the distributions of the input quantities are Gaussian.

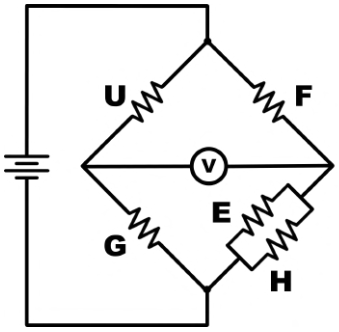


Histogram of 10^6 replicates of the value of V simulated using the Monte Carlo method, and probability density (red curve) of the Gaussian distribution with the same mean and standard deviation as those replicates.

Wheatstone Bridge

The Wheatstone bridge is an electrical circuit used to obtain accurate measurements of resistance by balancing both sides of a bridge circuit, one of which includes the component with unknown resistance (resistor U). In its simplest version, the Wheatstone bridge comprises a DC power supply, a voltmeter, and four resistors, one of which has adjustable resistance. The bridge illustrated here comprises three adjustable resistors, two of which are arranged in parallel so as to achieve finer control over their joint resistance, which is the harmonic mean of their individual resistances, R_E and R_H:

$$R_\mathrm{EH} = \frac{1}{R_\mathrm{E}^{-1} + R_\mathrm{H}^{-1}}$$



Wheatstone bridge comprising the resistor U whose resistance, R_U, one intends to measure, a resistor F with fixed resistance, and three resistors (G, E, and H) with adjustable resistances.

Resistor G is a General Radio decade resistor that can take values of resistance up to $1\,\mathrm{M}\Omega$ in steps of $0.1\,\Omega$, with relative standard uncertainty 0.05 %. Resistor E is an EICO decade resistor that can take values up to $100\,\mathrm{k}\Omega$ in steps of $1\,\Omega$, with relative standard uncertainty 0.5 %, and resistor H is a Heathkit RS-1 Resistance Substitution Box that allows the user to select one of several values of resistance.

The choice of instrumentation pays homage to a bygone era of analog electrical devices. The General Radio Company designed and manufactured test equipment for resistance, inductance, and capacitance, from 1915 until 2001, in West Concord MA. The Electronic Instrument Company (EICO) was established in Brooklyn NY, in 1945, and remained in business for 54 years. Besides test equipment, EICO also produced Geiger counters, amateur radio, and high-fidelity audio equipment.

We assume that the measurement experiment was carried out quickly enough, and at sufficiently low voltage (4 V), so that changes in resistance caused by heating of the resistors are negligible. We also assume that the error is negligible in achieving zero volt when balancing the bridge by adjusting the resistances of G, E, and H, thus reaching the point where $R_\mathrm{U}/R_\mathrm{G} = R_\mathrm{F}/R_\mathrm{EH}$. Hence, we have the

following measurement equation for R_U:

$$R_U = \frac{R_G R_F}{R_{EH}} = R_G R_F \left(R_E^{-1} + R_H^{-1} \right)$$

The observed resistance values with the associated standard uncertainties are listed in the table alongside. Since R_U is not a simple product of powers of R_E, R_F, R_G, and R_H, the approximation used above, for the uncertainty of the volume of the storage tank, cannot be used here. For this we use the Gauss method in its general form, which relates the uncertainties associated with uncorrelated input quantities R_E, R_F, R_G, and R_H, with the output quantity R_U:

$$u^2(R_U) \approx \left(\frac{\partial R_U}{\partial R_E}\right)^2 u^2(R_E) + \left(\frac{\partial R_U}{\partial R_F}\right)^2 u^2(R_F) + \left(\frac{\partial R_U}{\partial R_G}\right)^2 u^2(R_G) + \left(\frac{\partial R_U}{\partial R_H}\right)^2 u^2(R_H).$$

The partial derivatives of the measurement model are given in the table alongside. By substituting these partial derivatives into the expression above, we obtain

$$u^2(R_U) = \frac{R_G^2 R_F^2}{R_E^4} u^2(R_E) + \frac{R_U^2}{R_F^2} u^2(R_F) + \frac{R_U^2}{R_G^2} u^2(R_G) + \frac{R_G^2 R_F^2}{R_H^4} u^2(R_H).$$

Finally, the estimate of the measurand is

$$R_U = 909\,\Omega \times 997\,\Omega \times \left(\frac{1}{951\,\Omega} + \frac{1}{225.2\,\text{k}\Omega} \right) = 957\,\Omega,$$

with associated standard uncertainty $u(R_U) \approx 7\,\Omega$.

The *NIST Uncertainty Machine*[7] can produce the results in a single stroke. Modeling all the resistances as Gaussian random variables with means equal to the observed values and standard deviations equal to the standard uncertainties, we obtain not only $R_U = 957\,\Omega$ and $u(R_U) = 7\,\Omega$, but also a probability distribution for R_U and, in turn, a 95 % coverage interval for the true value of R_U, which ranges from 943 Ω to 971 Ω. Furthermore, we learn that the (squared) uncertainties of the different resistances contribute to the $u^2(R_U)$ in these proportions: F, 48 %; G, 0.6 %; E, 52 %; and H 0.004 %.

RESISTOR	R	$u(R)$
E	951 Ω	5 Ω
F	997 Ω	5 Ω
G	909 Ω	0.5 Ω
H	225.2 kΩ	2.3 kΩ

Observed resistance values that result in zero volt potential difference across the Wheatstone bridge.

DERIVATIVE	VALUE
$\partial R_U/\partial R_E$	$-R_G R_F/R_E^2$
$\partial R_U/\partial R_F$	$R_G(R_E^{-1} + R_H^{-1}) = R_U/R_F$
$\partial R_U/\partial R_G$	$R_F(R_E^{-1} + R_H^{-1}) = R_U/R_G$
$\partial R_U/\partial R_H$	$-R_G R_F/R_H^2$

Partial derivatives of the output quantity, R_U, with respect to all four input quantities. These and other derivatives can be readily obtained using a variety of online tools such as www.wolframalpha.com

[7] T. Lafarge and A. Possolo. The NIST Uncertainty Machine. *NCSLI Measure Journal of Measurement Science*, 10(3):20–27, September 2015. doi:10.1080/19315775.2015.11721732

Resistance is a positive quantity while the Gaussian uncertainty model entertains the possibility of negative values. For this reason, the lognormal model is sometimes chosen. The Gaussian and lognormal models are just about identical when the relative uncertainties are small (say, < 5 %).

The measurement model considered above does not recognize the uncertainty associated with balancing the Wheatstone bridge. A more elaborate model that accounts for this is as follows:[8]

$$R_{\mathrm{U}} = \frac{U_0 R_{\mathrm{G}}(R_{\mathrm{F}} + R_{\mathrm{EH}})}{U_0 R_{\mathrm{EH}} + U(R_{\mathrm{F}} + R_{\mathrm{EH}})} - R_{\mathrm{G}}.$$

Here, U_0 is the potential difference across the terminals of the DC power supply, $U_0 = 4\,\mathrm{V}$, and U is the potential across the balanced bridge ($U \approx 0\,\mathrm{V}$). Uncertainty analysis of this more complete measurement model using the *NIST Uncertainty Machine* reveals that balancing the Wheatstone bridge becomes the dominant source of uncertainty of R_{U} if the uncertainty associated with U is larger than $5\,\mathrm{mV}$.

[8] H. Zangl, M. Zine-Zine, and K. Hoermaier. Utilization of software tools for uncertainty calculation in measurement science education. *Journal of Physics: Conference Series*, 588:012054, 2015. doi:10.1088/1742-6596/588/1/012054

Counting

Fuentes-Arderiu and Dot-Bach [2009, Table 1] report results of classifying and counting white blood cells (leukocytes) of different types in a blood smear, known as a differential leukocyte count. The typical procedure when such counting is done manually while examining the sample under the microscope, is to count 100 leukocytes in total, while keeping a tally of the different types of leukocytes.

In this case, there were 4 eosinophils among the 100 leukocytes that were counted. It is to be expected that, if another blood smear from the same patient were to be similarly examined, the number of eosinophils would turn out different from 4, owing to the vagaries of sampling.

This source of uncertainty is often modeled using either the binomial or the Poisson probability distributions. Since the probability of finding an eosinophil is small, these two models lead essentially to the same evaluation of this uncertainty component: that the proportion of eosinophils should vary by about $\sqrt{4}/100 = 2\,\%$ around the measured value of 4, which is taken as the estimate of the Poisson mean, whence the count will have standard deviation $\sqrt{4}$.

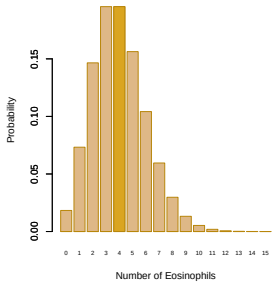
Counting the eosinophils involves: (i) identifying them, that is, defining the subset of the 100 leukocytes under examination that are eosinophils; (ii) actually counting the eosinophils that were identified; and (iii) qualifying the count with an evaluation of uncertainty, which should include contributions from sampling variability and from differences between examiners (which express identification uncertainty).

The standard for the identification task (i) should be the *holotype* (paradigm, reference exemplar) for an eosinophil. For species of plants and animals, the holotype is the individual specimen used to define a species, but there are no formal holotypes for different types of leukocytes. Because eosinophils are not identical copies of one another, accurate identification requires familiarity with their natural variability and reliance on distinctive traits that allow distinguishing them from the other types of leukocytes. For this reason, when different examiners count the same set of 100 leukocytes, it is likely that they will arrive at different counts for the different types of leukocytes.

Fuentes-Arderiu et al. [2007] have evaluated this source of uncertainty that is attributable to the effect of examiners, concluding that the coefficient of variation for the proportion of eosinophils was 69 %. Therefore, the uncertainty component for the count of eosinophils that arises from differences between examiners amounts to $4 \times 69\,\% = 3$ eosinophils.

LEUKOCYTES	n	$u_S(n)$	$u_B(n)$
Neutrophils	63	5	4
Lymphocytes	18	4	6
Monocytes	8	3	4
Eosinophils	4	2	3
Basophils	1	1	3
Myelocytes	1	1	1
Metamyelocytes	5	2	4

Table showing the leukocyte count (n). $u_S(n)$ quantifies the uncertainty attributable to sampling variability, and $u_B(n)$ does the same for differences between examiners.



Probabilities from Poisson distribution with mean 4, which is the number of eosinophils in the differential leukocyte count listed above.

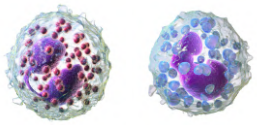
Holotype of a female *Agrias amydon phalcidon* butterfly from Brazil — Wikimedia Commons (Notafly, 2011).

The standard for the counting task (ii) is the unique finite set I comprising consecutive, positive integers, starting with 1, that can be put in one-to-one correspondence with the leukocytes that have been identified as being eosinophils: the measured value of the number of eosinophils is the largest integer in I. Task (ii) is counting *sensu stricto*, after identification, and is susceptible to counting errors. However, and in this case, since the numbers of leukocytes of the different types all are fairly small, and typically they are tallied using mechanical counters, we will assume that there are no counting errors above and beyond any identification errors.

Regarding task (iii), uncertainty evaluation, we need to take into account the fact that the total number of leukocytes that are identified and counted is fixed. Therefore, and for example, if an eosinophil is misclassified as a basophil, then the undercount for eosinophils results in an overcount for basophils. This means that the uncertainty evaluation for the counts cannot be performed separately for the different types of leukocytes, but must be done for all jointly, taking the effect of the fixed total into account: the so-called *closure constraint*.[9]

Performing a differential leukocyte count is equivalent to placing 100 balls (representing the 100 leukocytes) into 7 bins (representing the different types of leukocytes considered in this case), where the probability of a ball landing in a particular box is equal to the true proportion of the corresponding type of leukocyte in the subject's blood.

The probability model often used to describe the uncertainty associated with the numbers of balls that actually end-up in the different bins is the multinomial probability distribution. This model also takes into account the fact that no count can be negative. For the eosinophils, considering both sampling and examiner sources of uncertainty, their true count is believed to lie between 0 and 10 with 95 % probability, using methods reviewed under *Counts*.



Eosinophils (LEFT) are leukocytes that fight parasitic infections and mediate allergic reactions. Basophils (RIGHT) control the response to allergens — Wikimedia Commons (BruceBlaus, 2017). Unless the blood smear being measured is stained to emphasize basophils, they may be confused with eosinophils.

[9] F. Chayes. On correlation between variables of constant sum. *Journal of Geophysical Research*, 65(12):4185–4193, 1960. doi:10.1029/JZ065i012p04185

Surveying

In 2019, non-irrigated pastureland in Kansas was valued at around \$4620 per hectare (1 ha = 10 000 m^2). A plot, shaped like an irregular heptagon on an essentially flat plain, is for sale with asking price \$206 000. The seller offered to provide coordinates of the vertices in triplicate, determined using a portable, consumer-grade GPS receiver.

The potential buyer insisted that the triplicates should be obtained in three separate surveys. In each survey, the vertices were visited in random order, and the GPS receiver was turned off after taking a reading at a vertex, and then turned on again upon arrival at the next vertex, so that it would reacquire satellites and determine the location afresh.

These are the questions the potential buyer wishes a surveyor will answer: (i) How to estimate the plot's area? (ii) How to evaluate the uncertainty surrounding this estimate? (iii) How may have the seller come up with that asking price? The reason for this last question is that some understanding of the origin of the asking price may be a valuable element when the potential buyer will make a decision about how much to offer.

To estimate the plot's area one may use the *Surveyor's Formula*.[10] However, before using it, one needs to decide how to combine the triplicate determinations of the location of each vertex. One possible way consists of averaging them. For example, the average easting for vertex A is $e(\mathrm{A})/\mathrm{m} = (826 + 821 + 848)/3 = 831.7$. Let $(e(\mathrm{A}), n(\mathrm{A}))$, $(e(\mathrm{B}), n(\mathrm{B}))$, …, $(e(\mathrm{G}), n(\mathrm{G}))$ denote the averages of the Cartesian coordinates (easting and northing) of the triplicates at each vertex of the polygon in counterclockwise order (A, B, …, G). These are the coordinates of the large (green) dots in the plot alongside. The area of the shaded polygon is $S = 41.3\,\mathrm{ha}$, and it was computed as follows:

$$S = \frac{1}{2}\left(\begin{vmatrix} e(\mathrm{A}) & e(\mathrm{B}) \\ n(\mathrm{A}) & n(\mathrm{B}) \end{vmatrix} + \begin{vmatrix} e(\mathrm{B}) & e(\mathrm{C}) \\ n(\mathrm{B}) & n(\mathrm{C}) \end{vmatrix} + \cdots + \begin{vmatrix} e(\mathrm{F}) & e(\mathrm{G}) \\ n(\mathrm{F}) & n(\mathrm{G}) \end{vmatrix} + \begin{vmatrix} e(\mathrm{G}) & e(\mathrm{A}) \\ n(\mathrm{G}) & n(\mathrm{A}) \end{vmatrix}\right),$$

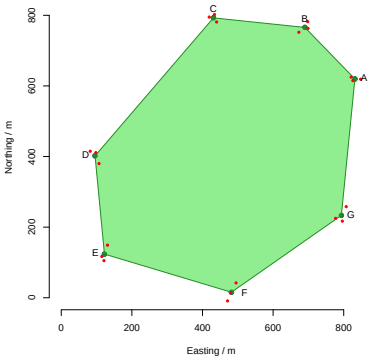
where $\left|\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right| = ad - bc$.

The question may well be asked of why the averages of the triplicates, instead of some other summary. The average will be optimal when the measurement errors affecting the easting and northing coordinates are independent and Gaussian, and the goal is to minimize the mean squared error of the estimates of the vertices.

Given the replicated determinations that were made of the locations of the vertices, it is possible to construct many different versions of the heptagon by choosing one of the three replicates made for vertex A, one of the three made for vertex B, etc. Each of these heptagons is consistent with the measurements that were made. Running through all $3^7 = 2187$ possible combinations of vertex determinations

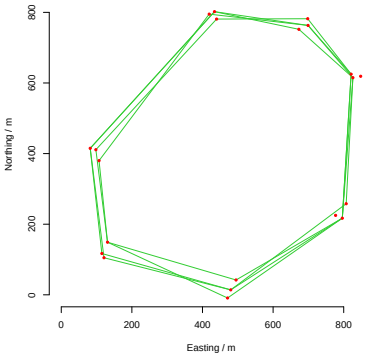
	EASTING / m			NORTHING / m		
A	826	821	848	615	625	619
B	673	698	699	752	782	763
C	440	419	434	781	795	802
D	82	98	107	415	411	380
E	131	121	115	149	105	117
F	471	495	480	-9	42	14
G	796	807	777	217	258	225

Coordinates of vertices of heptagonal plot of pastureland in Kansas, USA. One of the three determinations of location for vertex A has coordinates (826, 615), and similarly for all the others.



Plot of pastureland in Kansas, USA. The small (red) dots mark the triplicates of the vertices as determined by a GPS receiver, and the large (green) dots mark the averages of the triplicates.

[10] B. Braden. The surveyor's area formula. *The College Mathematics Journal*, 17 (4):326–337, 1986. doi:10.2307/2686282



Four of the $3^7 = 2187$ heptagons that can be constructed using the replicate determinations of the vertices.

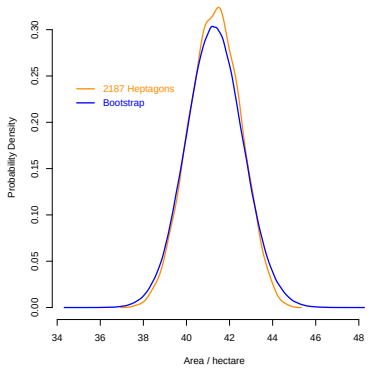
(each of which comprises a pair of values of easting and northing), and computing the areas of these alternative heptagons, yields a set of 2187 conceivable values for the area, whose average and median both equal 41.3 ha.

The area of the largest of these 2187 heptagons is 44.6 ha, with corresponding value 44.6 ha × \$4620/ha ≈ \$206 000, which explains the likely rationale behind the asking price. Since the area of the smallest heptagon is 37.6 ha, the same rationale would support an offer of 37.6 ha × \$4620/ha ≈ \$174 000.

However, an offer based on a value for the area close to the average area is more likely to be accepted by the seller than one that is as deviant from the average, but on the low side, as the one that the seller's asking price is based on, which is correspondingly on the high side. But the buyer should also take into account the uncertainty associated with the area.

Considering that each replicate of each vertex appears in $3^6 = 729$ heptagons built as just described, hence that there are correlations between the 2187 areas of the alternative heptagons, the standard deviation of these areas, 1.17 ha, may not be a reliable evaluation of the uncertainty associated with the area of the plot of land.

To evaluate this uncertainty, the buyer hires a statistician, whose first task is to quantify the uncertainty associated with the measurement of each vertex. The statistician applies the Fligner-Killeen test[11] to the replicated determinations of the easting and northing coordinates of the vertices of plot, and concludes that there is no reason to doubt that all 14 sets of replicates have the same variance. The statistician proceeds by pooling the variances of the 14 groups of replicates, which yields a standard uncertainty of 16 m (on 28 degrees of freedom) for an individual determination of the easting or northing of a vertex.



Probability density estimates for the area of the heptagon: based on the areas of the 2187 alternative heptagons, and on the bootstrap. The former (dark orange) ignores the correlations between the areas of the alternative polygons: the corresponding standard deviation is 1.17 ha. The latter (blue) reflects the impact of measurement errors affecting the easting and northing coordinates of each vertex, and recognizes the small numbers of replicates per vertex: it has heavier tails, and the corresponding standard deviation is 1.32 ha.

[11] M. A. Fligner and T. J. Killeen. Distribution-free two-sample tests for scale. *Journal of the American Statistical Association*, 71(353):210–213, March 1976. doi:10.2307/2285771

```
east = array(c(826, 673, 440, 82, 131, 471, 796, 821, 698, 419, 98,
               121, 495, 807, 848, 699, 434, 107, 115, 480, 777),
               dim=c(7,3))
north = array(c(615, 752, 781, 415, 149, -9, 217, 625, 782, 795,
                411, 105, 42, 258, 619, 763, 802, 380, 117, 14, 225),
               dim=c(7,3))
z = data.frame(east=c(east), north=c(north),
         east.vertex=I(paste0("E", rep(1:7, 3))),
         north.vertex=I(paste0("N", rep(1:7, 3))))
fligner.test(x=c(z$east, z$north), g=c(z$east.vertex, z$north.vertex))

east.s = apply(east, 1, sd)
north.s = apply(north, 1, sd)
s = sqrt(sum((3-1)*east.s^2 + (3-1)*north.s^2) /
       ((3-1)*length(east.s) + (3-1)*length(north.s)))
s.nu = (3-1)*length(east.s) + (3-1)*length(north.s)
c(s=s, s.nu=s.nu)
```

The pooled variance for easting and northing is the sum of the sums of squared deviations from their averages for the values of easting and northing, over all the vertices, divided by the sum of the corresponding numbers of degrees of freedom ($3 - 1$) per vertex. The pooled standard deviation, `s`, is the square root of the pooled variance.

The statistician's next task is to propagate this uncertainty to the uncertainty of the area, which she does employing the parametric statistical bootstrap [Efron and Tibshirani, 1993]. This involves repeating the following two steps a large number of times:

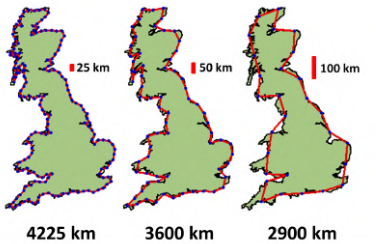
- For each vertex $i = 1, \ldots, 7$ in turn, simulate an easting of the form $e_i + \varepsilon_i$ and a northing of the form $n_i + \nu_i$, where (e_i, n_i) are the averages of the three determinations of easting and northing of vertex $i = 1, \ldots, 7$, and ε_i and ν_i represent measurement errors with zero mean and standard deviation 16 m — these measurement errors are drawings from Student's t distributions with 28 degrees of freedom, rescaled to have this standard deviation.
- Use the Surveyor's Formula to compute the area of the heptagon whose vertices' locations were simulated in the previous step.

The statistician repeated these steps one million times and found that the average of the areas of the simulated heptagons was the same as the area determined originally, and that the standard deviation of the simulated areas was 1.3 ha. In light of this fact, the statistician suggested to the buyer than an offer between $(41.3 - 1.3)\text{ha} \times \$4620/\text{ha} = \$184\,800$ and $(41.3 + 1.3)\text{ha} \times \$4620/\text{ha} = \$198\,612$ would be reasonable.

```
e = apply(east, 1, mean)
n = apply(north, 1, mean)
m = length(e)

K = 1e6
areaB = numeric(K)
for (k in 1:K)
{
  eB = e + s * rt(m, df=s.nu)/sqrt(s.nu/(s.nu-2))
  nB = n + s * rt(m, df=s.nu)/sqrt(s.nu/(s.nu-2))
  surv = (eB[m]*nB[1] - nB[m]*eB[1])
  for (i in 1:(m-1)) {
    surv = surv + (eB[i]*nB[i+1] - nB[i]*eB[i+1])}
  areaB[k] = (abs(surv)/2) / 10000
}
c(mean(areaB), sd(areaB))
```

The case just discussed involves a rather simple geometric figure: a heptagon whose boundary is clearly well defined. In practice, one often has to deal with more complex situations. Benoit Mandelbrot[12] famously asked the question "How long is the coast of Britain?" It so turns out that the answer to this question depends on the spatial scale at which the question is considered: or, in other words, on the size of the ruler used to measure it. Mandelbrot notes that "geographical curves are so involved in their detail that their lengths are often infinite or, rather, undefinable." In fact, the apparent length of the coastline decreases as the length of the ruler increases.

[12] B. Mandelbrot. How long is the coast of Britain? statistical self-similarity and fractional dimension. *Science*, 156:636–638, May 1967. doi:10.1126/science.156.3775.636



The estimated length of the UK coastline depends on the size of the ruler used (modified from Gurung [2017]).

Weighing

A laboratory weight C, of nominal mass 200 g, is to be calibrated using two previously calibrated reference weights A and B, whose masses exceed 200 g by 0.22 mg and 0.61 mg, respectively, both known to within 0.14 mg with 95 % probability, which suggests that these may be class E_2 weights.[13]

The calibration involves determining three mass differences using a mass comparator: the observed difference between the masses of A and B is $D_{AB} = -0.38\,\text{mg}$, and similarly $D_{AC} = -1.59\,\text{mg}$ and $D_{BC} = -1.22\,\text{mg}$.

Since the weight A has a nominal mass 200 g, we write $m_A = 200\,\text{g} + \delta_A$, where δ_A is the true deviation from the nominal mass, and similarly for the other weights. That is, we have the following simultaneous *observation equations*:[14]

$$\begin{aligned} D_{AB} &= \delta_A - \delta_B + \varepsilon_{AB} \\ D_{AC} &= \delta_A - \delta_C + \varepsilon_{AC} \\ D_{BC} &= \delta_B - \delta_C + \varepsilon_{BC} \end{aligned}$$

where ε_{AB}, ε_{AC}, and ε_{BC} denote the (non-observable) measurement errors incurred in the mass comparator. The conventional approach[15] involves finding values for δ_A, δ_B, and δ_C, that minimize the sum of the squared errors,

$$(\delta_A - \delta_B - D_{AB})^2 + (\delta_A - \delta_C - D_{AC})^2 + (\delta_B - \delta_C - D_{BC})^2,$$

subject to the constraint $\delta_A + \delta_B = 0.83\,\text{mg}$, which is one of several alternative constraints that could be applied.

The solution of this constrained linear least squares problem produces the estimate $\widehat{\delta_C} = 1.82\,\text{mg}$, with associated uncertainty $u(\widehat{\delta_C}) = 0.049\,\text{mg}$. Even though the maximum permissible error for a 200 mg class E_1 weight is 0.10 mg, it would be inappropriate to place the weight C into this class, considering that the calibrants are class E_2 weights.

Alternatively, an estimate of δ_C can be obtained using Bayesian statistical methods. For this, we model the measured mass differences probabilistically, as outcomes of Gaussian random variables:

$$\begin{aligned} D_{AB} &\sim \text{GAU}(\delta_A - \delta_B, \sigma), \\ D_{AC} &\sim \text{GAU}(\delta_A - \delta_C, \sigma), \\ D_{BC} &\sim \text{GAU}(\delta_B - \delta_C, \sigma). \end{aligned}$$

For example, the observed value of D_{AB} is viewed as a drawing from a Gaussian distribution with mean $\delta_A - \delta_B$ and standard deviation σ.

Radwag AK-4/2000 Automatic Mass Comparator (Radom, Poland).

[13] International Organization of Legal Metrology (OIML). *Weights of classes E_1, E_1, E_2, F_1, F_2, M_{1-2}, M_2, M_{2-3}, and M_3 — Part 1: Metrological and technical requirements*. Bureau International de Métrologie Légale (OIML), Paris, France, 2004. URL https://www.oiml.org/en/files/pdf_r/r111-1-e04.pdf. International Recommendation OIML R 111-1 Edition 2004 (E)

[14] P. E. Pontius and J. M. Cameron. *Realistic Uncertainties and the Mass Measurement Process — An Illustrated Review*. Number 103 in NBS Monograph Series. National Bureau of Standards, Washington, DC, 1967. URL http://nvlpubs.nist.gov/nistpubs/Legacy/MONO/nbsmonograph103.pdf

[15] R. N. Varner and R. C. Raybold. *National Bureau of Standards Mass Calibration Computer Software*. NIST Technical Note 1127. National Bureau of Standards, Washington, DC, July 1980. URL https://nvlpubs.nist.gov/nistpubs/Legacy/TN/nbstechnicalnote1127.pdf

We also use probability distributions to express what we know about the deviations from nominal of the masses of weights A and B, thus:

$$\delta_A \sim \mathrm{GAU}(0.22\ \mathrm{mg}, 0.07\ \mathrm{mg}),$$
$$\delta_B \sim \mathrm{GAU}(0.61\ \mathrm{mg}, 0.07\ \mathrm{mg}).$$

All we know about weight C is that it has a nominal mass of 200 g, but we also have good reasons to believe that its true mass lies within a reasonably narrow interval centered at 200 g. Providing a generous allowance for the length of this interval, we adopt the model

$$\delta_C \sim \mathrm{GAU}(0\ \mathrm{mg}, 100\ \mathrm{mg}).$$

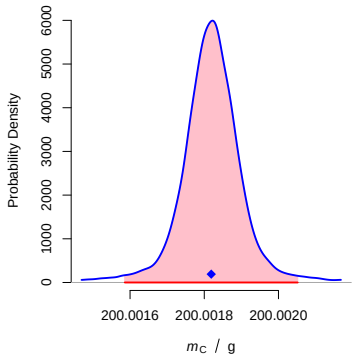
The fact that this prior standard deviation is comparable to the maximum permissible error for a class M_3 weight, does not signify that the weight C may be of this class. Rather, this choice serves only to give the data ample opportunity to make themselves heard, unencumbered by overly restrictive prior assumptions.

Since the Bayesian approach requires that all unknown parameters be modeled probabilistically, we need to assign a probability distribution also to the standard deviation, σ, of the measurement errors. To this end, we assume that the true value of σ is *a priori* equally likely to be larger or smaller than 1 mg, and assign a half-Cauchy distribution to σ, with median 1 mg. This choice provides great latitude for the value that σ may truly have, and gives the data ample opportunity to express themselves.

```
require(rstan); require(robustbase)
m = "data { real DAB; real DAC; real DBC; }
     parameters { real dA; real dB; real dC;
                  real<lower=0> sigma; }
     model {
       // Prior distributions
       dA ~ normal(0.22, 0.07); dB ~ normal(0.61, 0.07);
       dC ~ normal(0.00, 100);  sigma ~ cauchy(0.0, 1.0);
       // Likelihood
       DAB ~ normal(dA - dB, sigma); DAC ~ normal(dA - dC, sigma);
       DBC ~ normal(dB - dC, sigma); }"

fit = stan(model_code = m,
        data = list(DAB = -0.38, DAC = -1.59, DBC = -1.22),
        init = function() list(dA=0.22, dB=0.61, dC=1.8, sigma=0.1),
        warmup=75000, iter=250000, chains=4, cores=4, thin=25,
        control= list(adapt_delta=0.99999, max_treedepth=15))

dC.posterior = rstan::extract(fit)$dC
c(MEAN=huberM(dC.posterior)$mu, SD=Qn(dC.posterior))
```

The results of this Bayesian calibration are $m_C = 200\,001.82$ mg, give or take 0.23 mg, with 95 % probability.

The Monte Carlo Markov Chain method, implemented using the Stan modeling language in tandem with the R package `rstan` as detailed alongside, was used to draw a large sample from the posterior probability distribution of δ_C. A robust estimate of the mean of this sample equals 1.82 mg (which happens to be identical to the least squares estimate above), and a robust estimate of its standard deviation equals 0.074 mg, which is substantially larger than the uncertainty associated with the least squares estimate.



Probability density for m_C produced by the Bayesian calibration. Its mean value (indicated by a blue diamond), is the calibrated value of m_C. The horizontal, red line segment indicates an interval of half-length 0.23 mg that, with 95 % probability, is believed to include the true value of m_C.

Optimal design of experiments can use the results of uncertainty propagation as a guide. Consider a situation where we wish to determine the individual weights of three gold coins with the smallest uncertainty possible. We have access to a good balance but only for a limited time, enough to perform three weighings. The uncertainty associated with each weighing in this balance is constant, and does not depend on the mass being weighed, $u(m) = u$.

We could devise two experimental designs: (1) weigh each coin individually or (2) weigh them in pairs (coin 1 and coin 2 together, then coin 1 and coin 3 together, and finally coins 2 and 3 together). This is the measurement model corresponding to the latter design:

$$\begin{aligned} m_1 &= \tfrac{1}{2}\big(+ m_{1+3} + m_{1+2} - m_{2+3}\big), \\ m_2 &= \tfrac{1}{2}\big(- m_{1+3} + m_{1+2} + m_{2+3}\big), \\ m_3 &= \tfrac{1}{2}\big(+ m_{1+3} - m_{1+2} + m_{2+3}\big). \end{aligned}$$

Applying Gauss's formula to these expressions yields, for example,

$$\begin{aligned} u^2(m_1) &= \left(\frac{\partial m_1}{\partial m_{1+3}}\right)^2 u^2(m_{1+3}) \\ &\qquad + \left(\frac{\partial m_1}{\partial m_{1+2}}\right)^2 u^2(m_{1+2}) \\ &\qquad\qquad + \left(\frac{\partial m_1}{\partial m_{2+3}}\right)^2 u^2(m_{2+3}) \\ &= \frac{1}{4}u^2 + \frac{1}{4}u^2 + \frac{1}{4}u^2, \end{aligned}$$

and similarly for $u(m_2)$ and $u(m_3)$. Thus,

$$u(m_1) = u(m_2) = u(m_3) = u\sqrt{3/4}.$$

That is, by weighing the three coins in pairs we achieve 13 % lower uncertainty than by weighing them separately. Since the expressions above are linear combinations of the weighings, Gauss's formula is exact in this case.

Ranking

Ranking is assigning a place for an object being measured in an ordered sequence of standards, based on the value of a property whose values can be ordered from smallest to largest but not necessarily quantified. To distinguish harder and softer pencil leads, for example, pencil manufacturers rank pencils on a grading scale: from 9B (super black, very soft) to 9H (a gray scratch, very hard).

HARDNESS	MINERAL
1	talc
2	gypsum
3	calcite
4	fluorite
5	apatite
6	orthoclase
7	quartz
8	topaz
9	corundum
10	diamond

The minerals defining the Mohs hardness scale.

THE MOHS SCALE OF HARDNESS is determined by comparing a mineral specimen against a set of reference standards by means of a scratch test, whose results place it in the rank order of increasing hardness. The Mohs reference standards[16] are samples of various minerals with ordinal values 1 to 10 assigned to them without implying that the increase in hardness from gypsum to calcite is the same as the increase in hardness from apatite to orthoclase. For example, tourmaline typically scratches quartz and is scratched by topaz, hence its Mohs hardness is between 7 and 8. The numbers used to denote ranking order on an ordinal scale are nothing but labels for which arithmetic operations are not meaningful. Thus, numbers 1–10 could very well be replaced by letters A–J to convey the same message. In practice, when one says that the hardness of tourmaline is 7.5, all one means is that its hardness lies between the hardness of quartz and topaz.

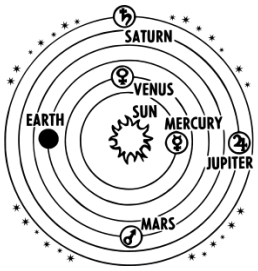
[16] C. Klein and B. Dutrow. *Manual of Mineral Science*. John Wiley & Sons, Hoboken, NJ, 23rd edition, 2007. ISBN 978-0-471-72157-4

Numbers are often used as labels with only an ordinal or nominal connotation. Examples of this use are the numbers used in the Saffir-Simpson ordinal scale of hurricane strength, and the numbers printed on the shirts of football players, where they serve to indicate the different players (nominal scale).

OUR ANCESTORS HAVE PONDERED FOR AGES the question of which planet is the closest to Earth. Most textbooks state that it is Venus because it makes the closest approach to Earth compared to any other planet.[17] The answer, however, depends on what is meant by "closest" — whether it means closest ever, closest on average, or closest most of the time —, because planets do not stand still and therefore distances between them are in constant flux.

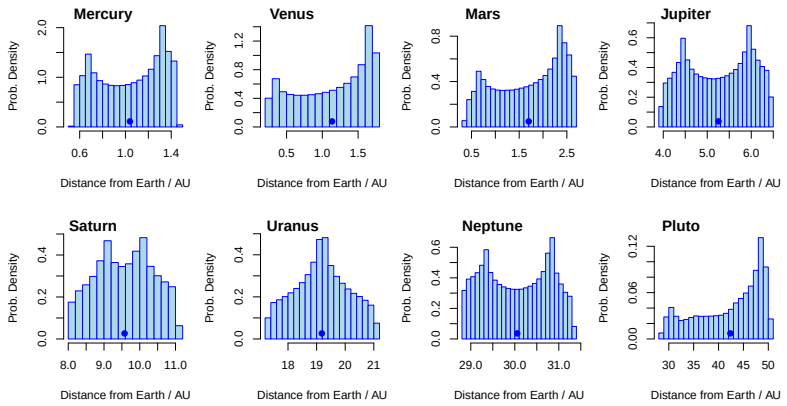
On January 1st, 2019, for example, Venus indeed was the planet closest to Earth, but that was no longer the case on the following February 24th, when Mercury moved closer. In the long term (over the period 2020-2420) Mercury will be Earth's closest neighbor 47 % of the time, Venus 37 % of the time, and Mars 16 % of the time, according to the NASA Jet Propulsion Laboratory HORIZONS system [Giorgini, 2015]. And it may be surprising that Pluto will be closer to Earth than Neptune 4 % of the time, even though its median distance to Earth is almost 1.5 times larger than Neptune's.

To characterize the positions of the planets relative to Earth properly, one needs to consider the distributions of the daily distances, as depicted in the histograms below. Except for Uranus, the average dis-



Which planet is closest to Earth? — Wikimedia Commons (Clon, 2016)

[17] T. Stockman, G. Monroe, and S. Cordner. Venus is not earth's closest neighbor. *Physics Today*, 72, March 2019. doi:10.1063/PT.6.3.20190312a



Histograms of the daily distances from Earth (expressed in astronomical units, AU), for the planets in the Solar System during the period 2020-2420. Each blue dot indicates the average distance from Earth.

tance does not represent a typical distance from Earth. Neither does the standard deviation of the daily distances capture the variability of the distances accurately. Even though the uncertainty of the distance from Earth to any other planet, computed for a particular day by the HORIZONS system, is rather small, the variability of the distances is quite large, and it is best communicated by means of the probability distributions depicted in these histograms, which may be interpreted as representing the uncertainty about the distance on a randomly selected day.

IN THE 2018 WINTER OLYMPICS, the gold, silver, and bronze medals in Ladies Single Skating were awarded to Alina Zagitova, Evgenia Medvedeva, and Kaetlyn Osmond, respectively, who earned total scores of 239.57, 238.26, and 231.02 points, from a panel of nine judges.

The medals are awarded considering only the final ranking of the athletes, regardless of whether the differences in the underlying scores are large or small. In 2018, a mere 1.31 point gap (that is, 0.55 %) separated Olympic gold from silver. How significant may this difference be considering the uncertainty that inevitably is associated with the assignment of scores?

Figure skating scores are produced by a complex scoring system that involves intrinsic levels of difficulty for technical elements, a priori weights, subjective evaluations made by nine judges independently of one another, and consideration of whether the elements are performed early or late during each routine. This example serves to illustrate how Monte Carlo methods — that is, methods based on simulations of contributions from recognized sources of uncertainty — can be used to carry out uncertainty evaluations. In this case, the Monte Carlo method will serve to shed light on the significance of the difference in scores that earned Zagitova the gold

Alina Zagitova performing in the ladies' skating short program at the 2018 Winter Olympics — Wikimedia Commons (David W. Carmichael, 2018).

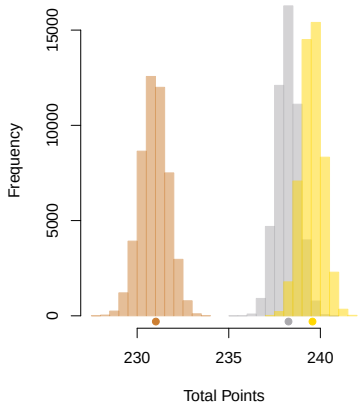
medal, and Medvedeva the silver. The table shows an excerpt from the score sheet for Zagitova's free skating component in the ladies finals: each executed technical element i has a particular, agreed-upon base value, b_i, and the quality of its execution is evaluated by nine judges. After removing the lowest and highest scores, the average score of the other seven is computed (trimmed mean) and added to the base value after multiplication by a predetermined weight, w_i. The combined score for element i is computed as follows, where $J_{i,j}$ denotes the score that judge j gave the athlete for the execution of this element:

$$s_i = b_i + \frac{w_i}{9-2}\left(\sum_{j=1}^{9} J_{i,j} - \min_{j=1,\ldots,9}\{J_{i,j}\} - \max_{j=1,\ldots,9}\{J_{i,j}\}\right)$$

The final scores are the sums of such element-specific scores, and certainly include expressions of the subjective, professional opinions of the nine judges. Given that judges do not always agree on their scores, it is reasonable to explore the extent of their disagreement. One way to assess the reliability of the judging scores is to simulate samples by randomly drawing scores, with replacement, from the set of actually observed scores, and then calculating the total score for each such random sample. This method is known as nonparametric bootstrap resampling [Efron and Tibshirani, 1993] and is widely used for uncertainty evaluations in science, medicine, and engineering. In this case, we generated 50 000 bootstrap samples, which enabled us to conclude that the probability of Medvedeva having won the gold medal was 6 %, thus quantifying the effect that judging uncertainty had upon the final result.

	EXECUTED ELEMENT		
JUDGE	3S	3F	2A
1	2	3	1
2	2	2	1
3	3	3	1
4	2	3	2
5	1	3	2
6	2	2	1
7	2	2	1
8	2	2	1
9	2	2	1
Base value, b	4.84	5.83	3.63
Weight, w	0.7	0.7	0.5
Total, s	6.24	7.53	4.20

Excerpt of score sheet for Alina Zagitova's final free skating component. 3S stands for triple Salchow, 3F for triple flip, and 2A for double Axel.



Probabilistic interpretation of the ladies single figure skating medal scores at the 2018 Winter Olympics.

Comparing

One of the most important applications of uncertainty evaluation is to compare two quantities whose measured values are surrounded by uncertainty. There is no margin for doubt when comparing numbers about which there is no uncertainty: everyone agrees that $9 > 7$. But it is impossible to decide conclusively whether meitnerium-277 and meitnerium-278 isotopes have the same or different longevity, considering that their half-lives[18] are estimated as $t_{1/2}(^{277}\mathrm{Mt}) = 9$ s and $t_{1/2}(^{278}\mathrm{Mt}) = 7$ s with standard uncertainties 6 s and 3 s, respectively. We shall illustrate five kinds of comparisons:

[18] G. Audi, F.G. Kondev, M. Wang, W. J. Huang, and S. Naimi. The NUBASE2016 evaluation of nuclear properties. *Chinese Physics C*, 41(3):030001–1–138, March 2017. doi:10.1088/1674-1137/41/3/030001

(i) a set of replicated observations of a quantity with the specified target value that the quantity is supposed to have;

(ii) a value measured by a user of a reference material with the corresponding certified value;

(iii) a set of replicated observations with a reference value qualified with an associated uncertainty;

(iv) two independent sets of observations obtained using two different methods for measuring the same quantity;

(v) contributions from different sources of uncertainty.

Comparing Replicated Determinations with Target Value

A particular kind of artillery shell is supposed to be loaded with 333 g of propellant. The values of the mass of propellant in 20 such shells, expressed in g, were: 295, 332, 336, 298, 300, 337, 307, 312, 301, 333, 344, 340, 339, 341, 297, 335, 345, 342, 322, 331.

The conventional treatment of this problem[19] involves computing the difference between the average of these 20 determinations, 324 g, and the specified target value, using as unit the standard uncertainty of the average:

[19] M. G. Natrella. *Experimental Statistics*. National Bureau of Standards, Washington, D.C., 1963. National Bureau of Standards Handbook 91

$$\frac{324\,\mathrm{g} - 333\,\mathrm{g}}{18.3\,\mathrm{g}/\sqrt{20}} = -2.2.$$

The denominator has the standard deviation of the determinations, 18.3 g, divided by the square root of their number, which is the Type A evaluation of standard uncertainty for the average, according to the GUM (4.2.3). Therefore, the average of these determinations is 2.2 standard uncertainties below the specified target value.

Still according to the conventional treatment, this standardized difference is to be interpreted by reference to a Student's t distribution with 19 degrees of freedom. The probability that such random variable will take a value that is more than 2.2 units away from zero, in

either direction, is 4 %. The reason why we consider deviations from zero in either direction is that we are testing a difference between the mean of the measured values and the specified value, regardless of whether that mean is larger or smaller than this specified value.

That probability, 4 %, is called the p-value of the test. It is the probability of observing a difference at least as large, in absolute value, as the difference that was observed, owing to the vagaries of sampling alone, on the assumption that in fact there is no difference. For this reason, a small p-value is usually interpreted as suggesting that the observed difference is significant.

The test just described is a procedure for statistical inference: the derivation of a conclusion from a sample, where the confidence in the conclusion is characterized probabilistically. The validity of the results of all such procedures hinges on the adequacy of the model and on particular assumptions, which are much too often neglected or taken for granted.

In this case, the assumptions are that the values in the sample are like outcomes of independent, Gaussian random variables, all with the same mean and standard deviation. The *Probability* Appendix points out that independence is a powerful property and a costly assumption, which is next to impossible to verify empirically in most cases. However, the assumption that the data originate in a Gaussian distribution can be evaluated using the Anderson-Darling test,[20] for example. This test yields a p-value of 0.2 %, computed in R as

```
m = c(295, 297, 298, 300, 301, 307, 312, 322, 331, 332,
      333, 335, 336, 337, 339, 340, 341, 342, 344, 345
library(nortest); ad.test(m)$p.value
```

This suggests that the test aforementioned may not be appropriate for these data, and that conformity with the target value ought best be evaluated in some other way.

Wilcoxon's one-sample signed rank test[21] does not require that the distribution the data come from be Gaussian, only that it be symmetric. The corresponding p-value is 0.22, obtained in R as `wilcox.test(m, mu=333)$p.value`. Therefore, the result of this test contradicts the result of Student's t test above, suggesting that the observations are consistent with the target value.

This example shows that conclusions drawn from data depend on assumptions and models used to describe particular patterns of variability of the data, and that the conclusions may change drastically when assumptions or models change.

The p-value of a two-sided Student's t test can be calculated using a variety of software. Since any software may suffer from errors, it is recommended that important calculations be replicated using implementations developed independently of one another in different software environments

	command
R	`2*pt(-2.2, df=19)`
Python	`2*stats.t.cdf(-2.2, df=19)`
Excel	`t.dist.2t(2.2, 19)`

[20] 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. doi:10.1214/aoms/1177729437

[21] M. Hollander, D. A. Wolfe, and E. Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons, Hoboken, NJ, 3rd edition, 2014

In 2014, 29 teams of researchers were asked to analyze the same data about red cards in soccer, using statistical procedures of their choice. Twenty teams concluded that there is a significant correlation between a player's skin color and his being given a red card, whereas nine teams concluded that there is none [Silberzahn and Uhlmann, 2015]

Comparing Measured Value with Reference Value

When comparing two estimates of the same quantity, in particular a measured value and a certified value, while taking their uncertainties into account, the overlap of corresponding coverage intervals is not sufficient reason to conclude that the corresponding true values are identical [Possolo, 2020, Example 7.2.A].

Note that equality to within specified uncertainties is not a transitive relation. Thus, if objects A and B are found to have identical masses to within their uncertainties, and if the same is true for objects B and C, it does not necessarily follow that the masses of A and C also are identical to within their respective uncertainties.

The certified mass fraction of nickel in NIST Standard Reference Material (SRM) 59a ferrosilicon is 328 mg/kg with expanded uncertainty 73 mg/kg for 95 % coverage. This means that the corresponding true value lies between 255 mg/kg and 401 mg/kg with 95 % probability. Suppose that a user of this material measured the mass fraction of nickel and obtained 172 mg/kg with expanded uncertainty 132 mg/kg, also for 95 % coverage. Since the corresponding coverage interval, ranging from 40 mg/kg to 304 mg/kg, overlaps the interval above, the inference may be drawn that there is no significant difference between the true mean of the user's measurement and the true value of the measurand.

The difference between the two measured values is 328 mg/kg − 172 mg/kg = 156 mg/kg and the standard uncertainty of the difference between these values is the square root of the sum of the individual, squared standard uncertainties,

$$\sqrt{((73/2)\ \text{mg/kg})^2 + (132/2)\ \text{mg/kg})^2} = 75\ \text{mg/kg}.$$

This statistical test assumes that the two values being compared are outcomes of independent Gaussian random variables, and that their associated standard uncertainties are based on infinitely many degrees of freedom. The p-value is the probability of observing a difference as large or larger (in absolute value) than the difference that was observed, by chance alone, owing to the vagaries of sampling and measuring the material, if the corresponding true values were identical. A small p-value suggests a significant difference.

The test statistic is the standardized difference, $156/75 = 2.08$. The p-value of the test is the probability of a Gaussian random variable with mean 0 and standard deviation 1 being either smaller than -2.08 or larger than $+2.08$. This probability is 3.75 %, which in most cases suggests a significant difference.

Comparing Replicated Determinations with Reference Value

To validate a measurement method, a laboratory often makes measurements of a reference material, and then compares the measurement results with the certified value. NIST SRM 1944 is a mixture of marine sediments collected near urban areas in New York and New Jersey, intended for use in evaluations of analytical methods for the determination of polychlorinated biphenyls (PCB) and several other hydrocarbons in similar matrices.

A quality control test yielded the following replicates for the mass fraction of PCB 95: 63.9 µg/kg, 48.4 µg/kg, and 46.1 µg/kg. Their average and standard deviation are 52.8 µg/kg and 9.7 µg/kg. The Type A evaluation of the standard uncertainty associated with the average is $9.7\ \mu\text{g/kg}/\sqrt{3} = 5.6\ \mu\text{g/kg}$, on 2 degrees of freedom.

The certified mass fraction of PCB 95 in SRM 1944 is 65.0 µg/kg, with standard uncertainty 4.45 µg/kg. The comparison criterion is

$$t = \frac{52.8 - 65.0}{\sqrt{5.6^2 + 4.45^2}} = -1.7.$$

On the hypothesis of no difference between the mean of the laboratory results and the certified value, this should be approximately like an outcome of a Student's t random variable with effective number of degrees of freedom (ν) given by the Welch-Satterthwaite formula [JCGM 100:2008, G.4], where the infinity appearing in the denominator is the "effective" number of degrees of freedom associated with the uncertainty evaluation for the certified value:

$$\nu = \frac{(5.6^2 + 4.45^2)^2}{\frac{5.6^4}{2} + \frac{4.45^4}{\infty}} = 5.3.$$

The hypothesis of no difference between measured and certified values entails that the criterion t should be like an outcome from a Student's t-distribution with 5.3 degrees of freedom. The larger the absolute value of t is, the more surprising it is that it should have occurred by chance alone, without there actually being a difference between measured and certified values. The questionable "logic" behind conventional tests of hypotheses is that rare events should not happen. Here, however, the probability is 15 % that an absolute value of 1.7 or larger might happen by chance alone owing to the vagaries of sampling, a far cry from a rare event, hence the conclusion that there is insufficient reason to reject the hypothesis of equality between measured and certified values.

Since the probability is 15 % that such random variable will deviate from 0 by more than 1.7 standard deviations, we conclude that the laboratory measurements do not differ significantly from the certified value. This conclusion is contingent on the three replicated determinations the laboratory made being like a sample from a Gaussian distribution — an assumption that is next to impossible to verify reliably with so few observations. Still, the Shapiro-Wilk test of Gaussian shape, whose R implementation accommodates samples this small, yields a comforting p-value of 23 %.

Comparing Two Measurement Methods

Laboratory practice often involves comparing a new or less-established method with an established standard method. The mass concentration of fat in human milk may be determined based on the measurement of glycerol released by enzymatic hydrolysis of triglycerides [Lucas et al., 1987], or by the Gerber method [Badertscher et al., 2007], which measures the fat directly with a butyrometer, after separating the fat from the proteins.

The correlation coefficient for these two sets of measured values is quite high, 0.998, but it 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.

A paired t-test indicates that the mean difference does not differ significantly from zero.[22] 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

γ_{Trig}	γ_{G}	γ_{Trig}	γ_{G}	γ_{Trig}	γ_{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

Pairs of values of the mass concentration of fat in human milk (expressed in cg/mL) determined based on enzymatic hydrolysis of triglycerides (Trig), and by the Gerber method (G), from Bland and Altman [1999, Table 3].

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

fair chance that a statistical test will fail to detect a difference that is important in practice. And if they are of a large size, then a statistical test very likely will deem significant a difference that is irrelevant in practice.

For these reasons, Bland and Altman [1986] suggest that graphical methods may be particularly informative about the question of agreement between methods.

The Bland-Altman plot 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 similar to Tukey's mean-difference plot.[23]

In this case, the difference between the methods tends to be positive for small values of the measurand, and negative for large values. This feature can be illustrated using a variant of the Bland-Altman plot that recognizes such trend. Function `BA.plot` from R package `MethComp` was used to draw the Bland-Altman plots.

Two methods are commonly employed to obtain the linear equation that "converts" a value produced by the Gerber method into the value that Trig would be expected to produce: the so-called *Deming regression* and *Passing-Bablok regression*.

Deming regression fits a straight line to points of a scatterplot when both coordinates are measured with error (ordinary linear regression assumes that only the response variable is measured with error). Passing-Bablok regression estimates the coefficients a and b in

$$\gamma_{\text{Trig}} = a + b \times \gamma_{\text{G}}$$

as follows: the slope b is the median of the slopes of the straight lines between every pair of points (excluding any resulting slopes that are either 0 or infinity), and the intercept a is the median of the intercepts $\{y_i - bx_i\}$ determined by each of the points. In this case, these methods yield the following lines:

$$\begin{aligned} \text{Deming:} \quad & \gamma_{\text{Trig}} = 0.078 + 0.972 \times \gamma_{\text{Gerber}}, \\ \text{Passing-Bablok:} \quad & \gamma_{\text{Trig}} = 0.055 + 0.976 \times \gamma_{\text{Gerber}}. \end{aligned}$$

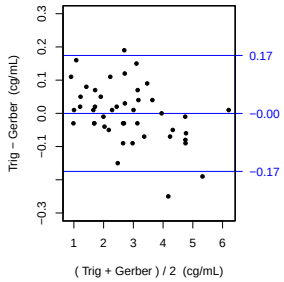
With 95 % confidence, the true slopes are believed to lie in these intervals:

$$\begin{aligned} \text{Deming Slope:} & [0.953, 0.988], \\ \text{Passing-Bablok Slope:} & [0.956, 0.995]. \end{aligned}$$

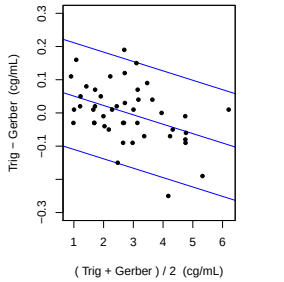
Since these intervals exclude the equivalence value of 1.000, we can conclude that the two methods do not provide equivalent results.

This article by Bland and Altman is the most often cited article in the *Lancet*, which reveals the exceptional interest that measurement issues enjoy in medicine. In 2014, *Nature* recognized this article as the 29th most-cited research of all time, over all fields.

[23] J. Chambers, W. Cleveland, B. Kleiner, and P. Tukey. *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA, 1983



Bland-Altman plot, with the average difference and the *limits of agreement* indicated by horizontal (blue) lines.



Bland-Altman plot recognizing that the differences between paired measured values depend on the averages of the same values.

These two regression lines can be computed using R functions defined in package `MethComp` [Carstensen et al., 2020] as follows:

```
Deming(x=Gerber, y=Trig, boot=TRUE)
PBreg(x=Gerber, y=Trig)
```

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.

To declare that two measurement methods are equivalent, not only should they produce results that are in agreement with due allowance for their respective uncertainties, over the relevant range of concentrations, but the measurement uncertainties that they typically achieve also should be in fair agreement.

Comparing Sources of Uncertainty

Assessing the homogeneity of a candidate reference material involves comparing the variability of the values of a property between units of the material, with their variability within units.

NIST SRM 2684c is a bituminous coal intended primarily for evaluations of analytical methods used for coals. Each unit of the material is a bottle containing 50 g of the finely powdered material. Between two and four aliquots from each of 23 selected bottles of the material were analyzed by X-ray fluorescence spectroscopy for aluminum content.

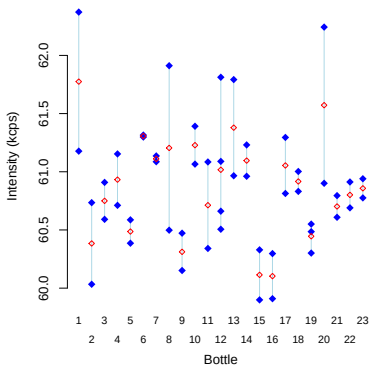
The conventional assessment of homogeneity is based on a statistical technique called *analysis of variance* (ANOVA).[24] Here, we will employ a model-based approach to evaluate potential heterogeneity, which is not observable directly but expresses itself in a parameter of the measurement model.

[24] R. A. Fisher. *Statistical Methods for Research Workers*. Hafner Publishing Company, New York, NY, 14th edition, 1973

The model, which will reappear in the discussion of *Consensus Building*, expresses the fluorescence intensity attributable to aluminum as

$$I_{ij} = \mu + \beta_j + \varepsilon_{ij},$$

where $j = 1, \ldots, n$ (with $n = 23$) denotes the bottle number, $i = 1, \ldots, m_j$ denotes the aliquot (subsample) from bottle j, μ is the overall mean intensity, β_j denotes the effect of bottle j on the measurement result, and ε_{ij} denotes the effect of aliquot i from bottle j. Only the $\{I_{ij}\}$ are observable.



X-ray fluorescence intensity from aluminum in aliquots drawn from bottles of NIST SRM 2684. Each red, open diamond represents the average of the determinations made in aliquots from the same bottle.

The bottle effects, $\{\beta_j\}$, are modeled as outcomes of random variables all with mean zero and standard deviation τ, and the aliquot effects $\{\varepsilon_{ij}\}$ are modeled as outcomes of random variables all with mean zero and standard deviation σ. These random variables do not need to be independent: it suffices that the bottle effects among themselves, and the aliquot effects among themselves, be *exchangeable*.

Material with undetectable inhomogeneity in the aluminum content corresponds to τ being zero: this means that readings of fluorescence intensity in aliquots from different bottles are *not* more variable than readings in aliquots from the same bottle. Suppose that $H = \theta(\mathbf{I})$ is a criterion that gauges heterogeneity (the opposite of homogeneity), where $\mathbf{I}$ denotes the set of 49 observations of fluores-

cence intensity, and θ denotes a function of these observations whose values are indications of heterogeneity. Suppose also that small values of H suggest that the material is homogeneous, and large values suggest that it is not.

Permute the elements of $\mathbf{I}$ randomly, similarly to how one would shuffle a deck of playing cards, so that the value a particular aliquot from a particular bottle may take the place of the value of any other aliquot, from any other bottle, the result being $\mathbf{I}^*$. If the material really is homogeneous, then $H^* = \theta(\mathbf{I}^*)$ should be fairly close to H.

Now, imagine repeating this process a large number K of times, thus obtaining $H_1^*, \ldots, H_K^*$, which together characterize the distribution of values of the heterogeneity criterion to be expected owing to the vagaries of sampling alone, on the assumption that the material indeed is homogeneous. Finally, compare the value of H that corresponds to the actual data, with the set $\{H_k^*\}$, and determine how "unusual" H may be among the $\{H_k^*\}$. If H should be unusually large, then this may warrant concluding that the material is heterogeneous.

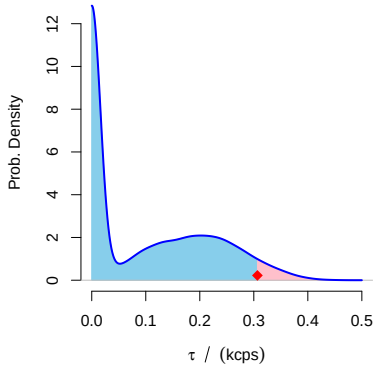
The criterion we shall use is an estimate of the standard deviation of the bottle effects, τ, which quantifies the component of variability above and beyond the within-bottle variability. There are many different ways of estimating τ, and it does not matter very much which one we will choose. For this example, we will rely on one of the most widely used estimators of τ — the restricted maximum likelihood estimator (REML).[25] We compute the value of τ corresponding to the measurement data (what above we called H), and also the values of τ for each of $K = 10\,000$ permutations of the data (what above we called $\{H_k^*\}$).

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

Out of 9990 permutations of the data (for 10 permutations the estimation procedure did not converge), only 458 yielded an estimate of τ that is larger than the estimate obtained for the actual data ($\tau = 0.31$ kcps). Therefore, the p-value of the permutation test of homogeneity is $458/9990 = 4.6$ %, which is commonly regarded as suggesting that the material is not homogeneous.

```
z = data.frame(
 bottle=c("B01", "B01", "B02", "B02", "B03", "B03", "B04", "B04", "B05",
   "B05", "B06", "B06", "B07", "B07", "B08", "B08", "B09", "B09", "B10",
   "B10", "B11", "B11", "B12", "B12", "B12", "B12", "B13", "B13", "B14",
   "B14", "B15", "B15", "B16", "B16", "B17", "B17", "B18", "B18", "B19",
   "B19", "B19", "B20", "B20", "B21", "B21", "B22", "B22", "B23", "B23"),
 kcps=c(62.37, 61.18, 60.73, 60.03, 60.91, 60.59, 60.71, 61.15, 60.39,
   60.59, 61.3, 61.32, 61.09, 61.14, 60.5, 61.91, 60.47, 60.15, 61.39,
   61.07, 61.08, 60.34, 60.51, 60.66, 61.81, 61.09, 61.79, 60.97, 60.96,
   61.23, 60.33, 59.9, 59.91, 60.3, 61.3, 60.81, 60.83, 61, 60.3, 60.49,
   60.55, 62.24, 60.9, 60.61, 60.8, 60.69, 60.91, 60.78, 60.94))
```

```
library(nlme)
z.lme = lme(kcps~1, random=~1|bottle, data=z, method="REML")
tau = as.numeric(VarCorr(z.lme)["(Intercept)", "StdDev"])

K = 10000
zB = z; tauB = rep(NA, K);
for (k in 1:K)
 { zB$kcps = sample(z$kcps, size=nrow(z), replace=FALSE)
   zB.lme = try(lme(kcps~1, random=~1|bottle, data=zB, method="REML"))
   if (class(zB.lme) == "try-error") {next
     } else {tauB[k] =
               as.numeric(VarCorr(zB.lme)["(Intercept)", "StdDev"])} }
tauB = tauB[complete.cases(tauB)]
```



Probability density of the estimates of τ obtained by permutations of the aluminum data (`tauB` in the alongside R code): the red diamond indicates the estimate of τ for the original data. The area shaded pink amounts to 0.46 % of the area under the curve.

The same model, $I_{ij} = \mu + \beta_j + \varepsilon_{ij}$, can also be fit to the data using a Bayesian procedure. The R package `brms`[26] provides a user friendly way to implement a wide variety of Bayesian regression models. This one-liner does it in this case:

```
brm(kcps ~ 1 + 1|bottle, data=z)
```

The best estimate of τ produced by this approach is $\tau = 0.28$ kcps. Since the true value of τ is believed to lie between 0.03 kcps and 0.53 kcps with 95 % probability, we can reject the hypothesis that $\tau = 0$ kcps and conclude confidently that there is evidence of heterogeneity in this material.

The first approach, based on permutations, involves fewer modeling assumptions than the Bayesian approach. However, all that it could do was perform a test of homogeneity, while the second approach both quantifies the heterogeneity and assesses its significance.

[26] P.-C. Bürkner. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411, 2018. doi:10.32614/RJ-2018-017

Calibrating

When a truck stops at a highway scale to be weighed, it applies a force to one or more load cells under the scale, which generates a potential difference between the electrical terminals that the load cells are connected to. Calibration is the procedure that establishes a relation between values of the force applied to a load cell and corresponding values of potential difference, thereby making possible to "translate" indications of voltage into values of force. These values of force, in turn, are translated into values of mass using the local value of the Earth's gravitational acceleration and Newton's second law of motion.

Calibrating a measuring instrument consists of determining a relationship between values of the measurand, and the typical, corresponding instrumental responses (or, *indications*), and characterizing the uncertainty surrounding such relationship. This is usually done by exposing the instrument to several different, known (up to measurement uncertainty) values of the measurand in measurement standards, making suitably replicated observations of the instrumental responses that these exposures generate, and finally deriving the typical responses from these observations.

The aforementioned relationship is often described by means of a *calibration function* that maps values of the measurand to typical (or, expected) values of the indications produced by the instrument being calibrated. For example, the result of calibrating a thermocouple for use as a thermometer is either a mathematical function that maps values of temperature into values of voltage, or a table that lists the values of voltage that correspond to specified values of temperature.

To be able to use the instrument to make measurements, the inverse relationship is needed, which produces an estimate of the value of the measurand given an observed instrumental response. This is variously called the *analysis function*, *measurement function*, or the *evaluation function*, depending on the field of application.

We begin by illustrating the development of calibration and analysis functions for the measurement of the mass concentration of chloromethane using gas chromatography and mass spectrometry, and in the process introduce criteria for model selection, and demonstrate Monte Carlo methods for uncertainty evaluation.

In this case, a very simple function, a cubic polynomial without the quadratic term, strikes just the right balance between goodness-of-fit to the calibration data and model simplicity. Many measurement systems, however, require calibration functions of much greater complexity.

For example, the calibration of capsule-type standard platinum resistance thermometers over the range 13.8033 K (triple point of hydrogen) to 273.16 K (triple point of water) in NIST SRM 1750 involved determining a polynomial of the 7th degree to describe the deviations between the ITS-90 reference curve for this range, and the actual values of resistance for these resistance thermometers [27]. An even more complex model is often used to characterize the dose-response of many bioassays, involving a five-parameter logistic function.[28]

One of the most complex calibration models used currently in science involves a Bayesian spline model with consideration of errors-in-variables that serves to convert measurements of carbon-14 concentration into measurements of the age of a biological material, in a technique known as *radiocarbon dating*.

[27] W. L. Tew and G. F. Strouse. *Standard Reference Material 1750: Standard Platinum Resistance Thermometers, 13.8033 K to 429.7485 K*. NIST Special Publication 260-139. National Institute of Standards and Technology, Gaithersburg, MD, November 2001. doi:10.6028/NIST.SP.260-139

[28] P. G. Gottschalk and J. R. Dunn. The five-parameter logistic: A characterization and comparison with the four-parameter logistic. *Analytical Biochemistry*, pages 54–65, 2005. doi:10.1016/j.ab.2005.04.035

Calibrating a GC-MS System

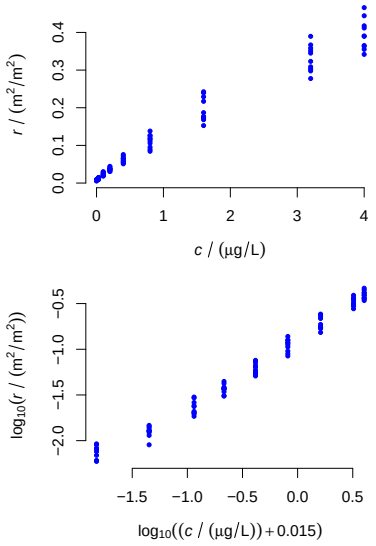
Chloromethane is a volatile organic compound with boiling point −24 °C at normal atmospheric pressure, and chemical formula CH_3Cl. It is currently used industrially as a reagent and solvent, and in the past was widely used as a refrigerant. Chloromethane is water-soluble and its concentration in water is usually measured using gas chromatography and mass spectrometry (GC-MS).[29]

[29] J. W. Munch. *Method 524.2. Measurement of Purgeable Organic Compounds in Water by Capillary Column Gas Chromatography/Mass Spectrometry*. National Exposure Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, OH, 1995. Revision 4.1

The table below lists replicated instrumental indications obtained with a GC-MS system to measure mass concentration of chloromethane, using fluorobenzene as internal standard [Lavagnini and Magno, 2007]: the indications are ratios between areas of peaks in the traces produced by the measuring system, one corresponding to chloromethane, the other corresponding to a known amount of the internal standard, which is injected into the system simultaneously with each sample of each chloromethane standard, thereby correcting for losses of the measurand (or, analyte) in the standard as it travels through the GC column.

Concentration of chloromethane, c (µg/L)								
0.00	0.03	0.10	0.20	0.40	0.80	1.60	3.20	4.00
0.009 22	0.012 87	0.024 12	0.036 82	0.051 04	0.111 98	0.174 22	0.344 97	0.355 10
0.009 10	0.012 68	0.020 21	0.038 46	0.053 50	0.084 40	0.172 28	0.297 68	0.341 71
0.006 91	0.014 31	0.020 90	0.031 08	0.064 27	0.095 43	0.168 29	0.308 67	0.365 22
0.008 31	0.012 29	0.020 33	0.036 36	0.055 83	0.118 92	0.152 62	0.277 52	0.363 19
0.007 60	0.009 01	0.023 62	0.044 51	0.071 74	0.125 51	0.229 08	0.351 52	0.417 58
0.009 01	0.011 42	0.019 58	0.037 59	0.057 60	0.089 32	0.216 99	0.302 68	0.389 76
0.006 06	0.014 70	0.026 16	0.030 71	0.075 69	0.116 85	0.186 97	0.389 64	0.411 68
0.008 03	0.013 76	0.018 47	0.034 26	0.066 60	0.138 12	0.176 93	0.323 14	0.390 48
0.005 93	0.012 90	0.030 00	0.037 08	0.059 65	0.126 42	0.242 47	0.358 24	0.465 81
0.006 03	0.012 80	0.029 38	0.042 27	0.064 50	0.105 84	0.239 47	0.366 87	0.444 20

A plot of the values of r against corresponding values of c shows that the dispersion of the replicated values of r increases substantially with increasing values of c. This undesirable feature is much reduced



Graphical representations of the data produced to calibrate a GC-MS system, before and after re-expression using logarithms. The choice of 0.015, which was added to $c/(\mu g/L)$ to avoid taking logarithms of zero, is arbitrary and inconsequential: in this case, it is half of the smallest positive value chosen for c in the calibration experiment.

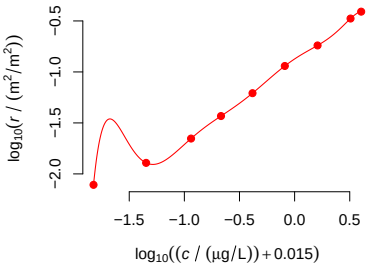
once the data are re-expressed using logarithmic scales, which also implies that the focus is on the relative uncertainties.

We will neglect the uncertainty surrounding the values of c because, in this particular case, in fact it is negligible by comparison with the dispersion of the replicated values of r. (Possolo [2015, E17] describes an instance of calibration where uncertainties surrounding the values of the measurand in the calibration standards, and the instrumental indications, both have to be recognized.)

For each of nine chloromethane calibration standards, ten replicate measurements of the ratio r of areas of peaks produced by the GC-MS measuring system, that correspond to chloromethane and to the internal standard [Lavagnini and Magno, 2007, Table 2].

MODEL SELECTION is the task of choosing a model to represent how $R = \log_{10}(r)$ varies as a function of $C = \log_{10}(c/(\mu g/L) + 0.015)$. Several polynomial models may be used to summarize the relationship between them: for example $R = \alpha + \beta C$, $R = \alpha + \beta_1 C + \beta_2 C^2 + \beta_3 C^3$, or $R = \alpha + \beta_1 C + \beta_3 C^3$, because one may either add or remove terms while searching for the best model. As more and more terms involving different powers of C are added to the model, the polynomial fits the data ever more closely. When to stop, and which model to choose?

Suppose we would summarize the replicated values of r that correspond to each value of c with their median, and fitted a polynomial of the 8th degree to these nine points. This polynomial fits the summary data exactly, but look how it behaves around the two leftmost points!



A polynomial may fit the data exactly and still be an awful calibration function.

While inappropriate here, polynomials of high degree are used occasionally as models. The International Temperature Scale ITS-90, for example, uses polynomials of the 9th and 15th order as reference functions.

THE GOAL IN MODEL BUILDING is to achieve a representation of the data that is accurate enough for the purpose the model is intended to serve, while keeping the model as parsimonious as possible. Parsimony, in this case, means small number of adjustable parameters, or low degree of the polynomial. The reason why parsimony matters is that simple models generally have better real-world performance than extravagant models, in the sense that they tend to be more accurate when applied to data similar to, but different from the data they were derived from.

For a polynomial model, fitting the model to the data amounts to finding values of the coefficients that make the graph of the polynomial pass as closely as possible to the data points. Several aspects of this issue are discussed under *Least Squares*.

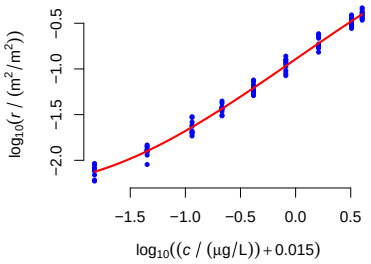
MODEL, φ	BIC(φ)
$\alpha + \beta_1 C$	−190
$\alpha + \beta_1 C + \beta_2 C^2$	−226
$\alpha + \beta_1 C + \beta_2 C^2 + \beta_3 C^3$	−231
$\boldsymbol{\alpha + \beta_1 C + \beta_3 C^3}$	**−235**
$\alpha + \beta_1 C + \beta_2 C^2 + \beta_3 C^3 + \beta_4 C^4$	−227
$\alpha + \beta_1 C + \beta_2 C^2 + \beta_3 C^3 + \beta_4 C^4 + \beta_5 C^5$	−222

The smaller the value of Bayesian Information Criterion, BIC, the more adequate the model for the data. In general, a difference in BIC values greater than 10 is strong evidence against the model with the higher BIC value, whereas a difference of less than 2 is considered insignificant. Thus, and in this case, the models in the third and fourth rows of this table are comparably adequate for the data.

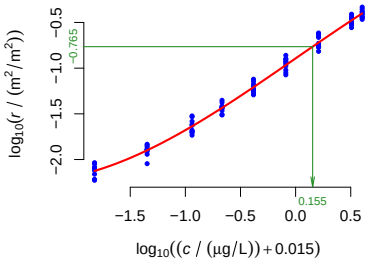
A RELIABLE GUIDE FOR MODEL BUILDING will strike a compromise between goodness of fit and simplicity. One such guide is the *Bayesian Information Criterion* (BIC) [Burnham and Anderson, 2004], which is explained under *Model Selection*. For now, it suffices to note that the smaller the BIC, the more adequate is the model for the data.

For the GC-MS calibration data listed above, the best model hap-

pens to be a polynomial of the third degree without the quadratic term, $\varphi(C) = \alpha + \beta_1 C + \beta_3 C^3$, with $\widehat{\alpha} = -0.8931$, $\widehat{\beta}_1 = 0.8327$, and $\widehat{\beta}_3 = -0.0473$. This defines the calibration function, which characterizes how the GC-MS measuring instrument responds when exposed to standard solutions of chloromethane.



Calibration function, whose graph is the red curve, is a polynomial of the third degree without the quadratic term.



Determination of the value of c that corresponds to an instrumental indication $r = 0.1718\,\mathrm{m^2/m^2}$. Inversion of the calibration function produces $\log_{10}((c/\mu\mathrm{g/L}) + 0.015) = 0.155$, hence $c = 1.41\,\mu\mathrm{g/L}$.

THE ANALYSIS FUNCTION is the mathematical inverse of the calibration function: ψ such that $\psi(\varphi(C)) = C$, for each value of C at which φ is defined. The analysis function is used to assign values of the measurand to samples whose mass concentration c of chloromethane is unknown, and which, upon injection into the GC-MS measuring instrument, produce a value of the ratio r.

Depending on the mathematical form of the calibration function φ, it may or may not be possible to derive an analytical expression (that is, a formula) for the analysis function ψ. However, it is always possible to determine it numerically given an observed value of R, by finding the values of C such that $\varphi(C) = R$. In case this equation is satisfied by more than one value of C, then some additional criterion needs to be employed to determine the appropriate solution: for example, the appropriate solution should lie between the minimum and maximum of the values of c in the standards used for calibration.

The mathematical inversion that leads from φ to ψ can be performed without any mathematics or computation at all: draw the graph of the calibration function φ on a sheet of transparent acetate, with the axis with values of c horizontal and increasing from left to right, and the axis with values of r vertical and increasing from bottom to top. Then flip the sheet and look at it from the back side, and rotate it so that the axis that was horizontal becomes vertical, and the one that was vertical becomes horizontal, the former now with values of c increasing from bottom to top, and the latter with values of r increasing from left to right. The resulting graph depicts the analysis function ψ.

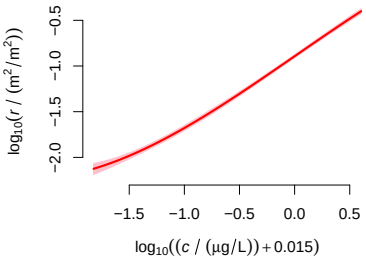
In this case the calibration function is a polynomial of the third degree, and indeed it is possible to solve $\varphi(C) = R$ analytically for C using a celebrated formula published in 1545 by Gerolamo Cardano, which implements the solution derived by Scipione del Ferro. In practice, however, even in cases like this, solving the equation numerically may be the more expeditious route, focusing most of the effort on determining the appropriate solution among the several that typically are available for polynomial calibration functions. And this is how the graph of ψ was constructed that is displayed alongside, by solving $\varphi(C) = R$ for C for many equispaced values of R.

The evaluation of the uncertainty surrounding the calibration and analysis functions may be performed using a Monte Carlo method, which in this case will be the non-parametric statistical bootstrap invented by Bradley Efron and explained to perfection by Diaconis and Efron [1983]. The uncertainty evaluation is based on the results from many repetitions of these two steps:

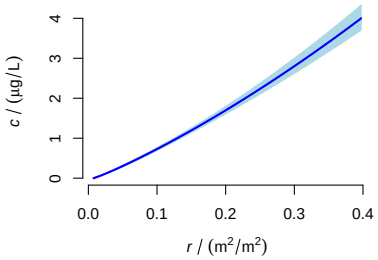
(1) Draw a sample of size 90 from the set of 90 pairs $\{(c_{ij}, r_{ij})\}$ listed in the foregoing table, uniformly at random, with replacement: this means that all pairs have the same probability of being selected, and that each pair may be selected more than once;

(2) Use this sample as if it were the original data, and select and build a calibration function as described above — this is called a *bootstrap replicate* of the calibration function.

Repeating these two steps 50 000 times, and finding the band that contains 95 % of the graphs of the resulting calibration functions, leads to the figure where the calibration curve is depicted in red with the pink uncertainty band. The similar, light blue band shown alongside, which surrounds the analysis function, was obtained by mathematical inversion of the upper and lower boundaries of the pink band that surrounds the calibration function.



Calibration function and 95 % coverage band derived from 50 000 bootstrap replicates of the calibration function.



Analysis function ψ and 95 % coverage band. Given an instrumental indication r, the function produces an estimate of the mass concentration c, and an evaluation of the associated uncertainty, in the form of a 95 % coverage interval

Shroud of Turin

The discovery of *radiocarbon dating* earned Willard F. Libby the 1960 Nobel Prize in Chemistry, and the accolade from the Nobel Committee that "seldom has a single discovery in chemistry had such an impact on the thinking in so many fields of human endeavor."

^{14}C atoms are continuously generated in the atmosphere as neutrons produced by cosmic rays strike nitrogen atoms, and eventually are absorbed by living organisms. The concentration of ^{14}C in the living tissues stays in equilibrium with its atmospheric counterpart until the organism dies. Thereafter, the ratio of concentrations of ^{14}C and of ^{12}C in the remains decreases steadily over time.

By measuring this ratio in the remains, and assuming that the ratio of concentrations of ^{14}C and ^{12}C in the atmosphere during the organism's lifetime was the same as it is today, it is possible to estimate how long ago the plant or animal died.

While simple in principle, radiocarbon dating is challenging in practice. First, the amount fraction of ^{14}C in pure carbon is minuscule: about 1 atom of ^{14}C per trillion atoms of carbon (of which the vast majority are ^{12}C and ^{13}C atoms). This implies that, in 4 grams

Caravaggio (1603-1604) *La Deposizione di Cristo*, Pinacoteca Vaticana, Vatican City — Wikimedia Commons (in the public domain, PD-US-expired).

of carbon, only one atom of ^{14}C will decay per second, on average. Therefore, radiocarbon dating based on measurements of activity requires fairly large samples of material. Mass spectrometry, which actually counts atoms of different mass numbers, has enabled radiocarbon dating of very small samples of material.

Second, radiocarbon dating rests on two key assumptions: (i) that the ratio of concentrations of ^{14}C and ^{12}C atoms in the atmosphere has remained constant over time, and equal to its present value; and (ii) that its value is the same for all biological tissues. Neither of these assumptions is valid. The first because the burning of fossil fuels (which contain no ^{14}C) has steadily decreased the fraction of ^{14}C in the atmosphere, while detonations of nuclear weapons from the 1940s until the early 1960s, increased it. The second because isotopic fractionation changes the relative concentrations of the three isotopes of carbon according to the provenance of the biological material used for dating.

These contingencies imply that accurate dating cannot be achieved without calibration, which establishes a correspondence between radiocarbon ages based on the ideal assumptions aforementioned, and known calendar ages of particular samples.

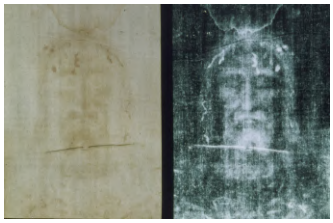
The most recent calibration curve is INTCAL20.[30] For the most recent 14 000 years, this curve is based entirely on tree-ring measurements, which can be dated by counting rings from outermost to innermost. Also, each ring's isotopic composition is a snapshot of the atmospheric composition at the time when the ring was growing.

The measurement of the age of the Shroud of Turin using radiocarbon dating is one of the most talked-about applications of the technique. The shroud is a linen cloth kept in the Cathedral of Saint John the Baptist, in Turin, Italy, which bears marks of the body of a tall, bearded man who may have been flogged. Some people believe that it is the burial cloth of Jesus of Nazareth.

Mass spectrometric measurements made in 1988 by Damon et al. [1989] at laboratories in Tucson (Arizona, U.S.A.), Oxford (England), and Zurich (Switzerland), yielded average radiocarbon age of 691 years Before Present (BP), with standard uncertainty 31 years.

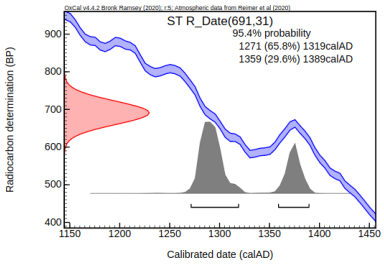
The resulting distribution of calendar age is a bizarre bimodal distribution whose mean (1317 AD) and standard deviation (40 years) tell us very little about the likely age of the shroud. Hence, it provides a cogent illustration of the fact that probability densities are well suited to capture the uncertainty of complex outcomes whereas summary estimates can be spectacularly deceiving. According to `OxCal`, the age of the Shroud lies between 1271 AD and 1319 AD with 65.8 % probability, and between 1359 AD and 1389 AD with 29.6 % probability, hence it is Medieval and not from Antiquity.

[30] T. J. Heaton, M. Blaauw, P. G. Blackwell, C. Bronk Ramsey, P. J. Reimer, and E. M. Scott. The INTCAL20 approach to radiocarbon calibration curve construction: a new methodology using Bayesian splines and errors-in-variables. *Radiocarbon*, pages 1–43, 2020. doi:10.1017/RDC.2020.46



Positive and negative versions of a portion of the Shroud of Turin — WikiMedia Commons (in the public domain, PD-US-expired).

By convention, radiocarbon ages are expressed as numbers of years before 1950, denoted as "before present" (BP) although there is some uncertainty about whether this means the the very beginning of 1950 [Townsley, 2017] or mid-1950 [Ramsey, 2009].



Calibration of the 1988 radiocarbon age measurement of the Shroud of Turin using the INTCAL20 [Reimer and et al., 2020] calibration curve, to obtain an estimate of the calendar age, as produced by the online version of `OxCal` v4.4.2 from the University of Oxford Radiocarbon Accelerator Unit [Ramsey, 2009].

Categorizing

Nominal and *ordinal* properties are kinds of *categorical* properties, which are qualitative [Agresti, 2019]. The values of a nominal property are names of sets of objects that have the same values of the (qualitative or quantitative) properties that define these sets. For example, when presented with an animal of the genus *Panthera*, one compares it with standard specimens of the five species in this genus, to determine whether the animal is a tiger, leopard, jaguar, lion, or snow leopard. This comparison may involve examining qualitative attributes such as the body shape, size, or color of the fur. If only a sample of tissue from the animal is available, then the comparison may involve sequencing particular areas of the genome, and comparing these sequences with paradigmatic sequences of known provenance that are available in gene databases.[31]

The values of ordinal properties can be ranked (ordered from smallest to largest) yet they are not quantitative: of them it can be said whether one is more or less than another, but not by how much. For example, the Mohs hardness of a mineral is determined by finding out which in a collection of reference minerals it scratches, and which it is scratched by.

[31] Y. Cho, L. Hu, and H. *et al.* Hou. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nature Communications*, 4:2433, September 2013. doi:10.1038/ncomms3433

Measuring Abortion Rates

Unsafe abortion caused 5 % to 13 % of maternal deaths worldwide during the period 2010–2014, and a large proportion of the abortions were performed unsafely.[32] The prevalence of abortion therefore is an important public health measurand. Having ever had an induced abortion is a nominal property of every woman, whose values are YES or NO. Determining its value reliably is challenging because women often are reluctant to report it.

[32] WHO. Preventing unsafe abortion. Evidence Brief WHO/RHR/19.21, World Health Organization, Geneva, Switzerland, 2019

In a randomized response, house-to-house survey conducted in Mexico City in 2001, each participating woman was asked one of two questions, selected at random, as if by tossing a fair coin: whether she had ever attempted to terminate a pregnancy, or whether she was born in the month of April.[33]

Only the woman being interviewed could see which of these two questions had been drawn for her, and she truthfully answered YES or NO to the question she was presented with. Since this survey technique preserves confidentiality, it tends to produce more reliable results than, for example, interviews where a woman is asked directly, face-to-face, the sensitive question about abortion.

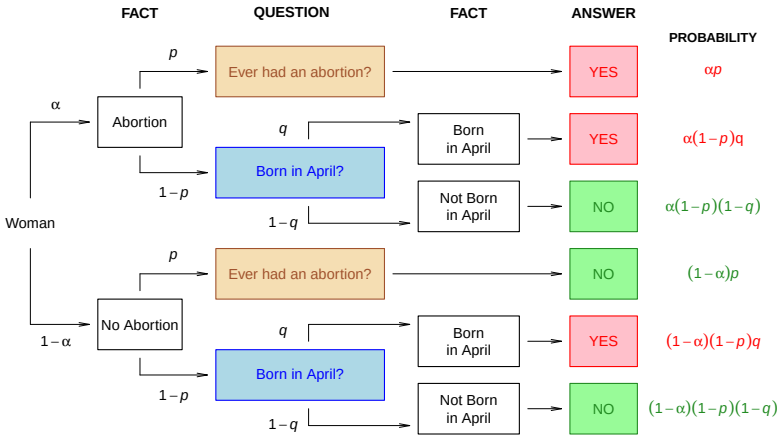
[33] D. Lara, J. Strickler, C. D. Olavarrieta, and C. Ellertson. Measuring induced abortion in Mexico: A comparison of four methodologies. *Sociological Methods & Research*, 32(4):529–558, May 2004. doi:10.1177/0049124103262685

This type of survey safeguards the confidentiality of responses and by doing so improves the reliability of its results. However, confidentiality could possibly be breached if the interviewer knew the participant personally, and also knew that she was not born in April. In such case, a YES answer reveals the attempted abortion.

Of the 250 women that participated in the house-to-house survey, 33 answered YES to the question they were presented with. This number includes women who had had an abortion and were asked the question about abortion, as well as women who were born in the month of April and were asked whether it was so, regardless of whether they had ever had an abortion. Since the survey design prevents determining individual values of the nominal property, the goal is to measure its prevalence, α, which is the proportion of

women who had ever attempted an abortion.

The following diagram shows how YES and NO answers may arise, where $p = 1/2$ is the probability of being asked the sensitive question, and $q = 1/12$ denotes the probability of having been born in April. The last column lists the probabilities of the different instances of YES and NO. Note that the six probabilities sum to 1.



Randomized response survey to measure prevalence of abortion.

The probability of YES is θ, which is the sum of the three terms above that appear in red in the last column of the diagram:

$$\theta = \alpha p + \alpha(1-p)q + (1-\alpha)(1-p)q = \alpha p + q(1-p).$$

Since the estimate of θ is $33/250$, $p = 1/2$ by design, and $q = 1/12$ on the assumption that births are equally likely to fall on any month of the year, α can be estimated by solving $33/250 = \alpha p + q(1-p)$ for α, which yields $\widehat{\alpha} = 271/1500 = 0.18$.

To evaluate the uncertainty, $u(\widehat{\alpha})$, let $\widehat{\theta}$ denote the estimate of θ, so that

$$\widehat{\alpha} = \widehat{\theta}/p - q(1-p)/p.$$

The second term on the right-hand side is a constant, whose variance therefore is zero. And the first term is a random variable divided by a constant. Now, the random variable, $\widehat{\theta}$, has a binomial distribution based on 250 trials, whose variance can be estimated as $\widehat{\theta}(1-\widehat{\theta})/250$, with $\widehat{\theta} = 33/250 = 0.132$. Therefore,

$$u(\widehat{\alpha}) = u(\widehat{\theta})/p = \sqrt{\frac{0.132(1-0.132)}{250}}/(1/2) = 0.043.$$

A 95 % coverage interval for α can be derived from a corresponding coverage interval for θ, which can be computed as described under *Counts*, finally to obtain $(0.10, 0.28)$, which is the output of the following R command:

```
(prop.test(x=33, n=250)$conf.int - q*(1-p))/p
```

The assumption that births are equally likely to fall in any of the twelve months of the year is approximately true for Mexico but is not quite true for the U.S. where the probability of a birth falling in April is only 0.079, while it is 0.091 for August.[34]

Considering that each value of q specifies one particular model for the randomized response survey, the uncertainty in q may be incorporated via *model-averaging*, and using the statistical bootstrap.

Assuming that q has a uniform distribution between 0.075 and 0.091 (the extreme rates of birthdays in the twelve months of the year, observed for the U.S.), the estimate of the prevalence of abortion becomes $\widetilde{\alpha} = 0.16$, and a 95 % uncertainty interval for the true value of α now ranges from 0.15 to 0.17.

[34] J. A. Martin, B. E. Hamilton, M. J. K. Osterman, and A. K. Driscoll. Births: Final data for 2018. National Vital Statistics Reports 68(13), National Center for Health Statistics, Centers for Disease Control and Prevention (CDC), Hyattsville, MD, November 2019

Uncertainty in Measurement Models

All measurements, however simple, are instances of model-based inference, and most measurements cannot be completed without some statistical data reduction because measurements are contingent on mathematical or statistical models. Building or selecting a measurement model is an integral part of measurement, and the same as other parts, typically it is surrounded by uncertainty.

Mass of Pluto

Modeling the motion of the heavenly bodies that comprise the solar system has fascinated scientists for centuries. As a feat of mathematical modeling and precision measurements, Neptune was discovered in 1846 based on the analysis of observational data about the motion of Uranus. This discovery remains one of the best examples of the power of the scientific method, and it prompted many at the time to look for the next planet that might lurk beyond the newly-discovered Neptune.

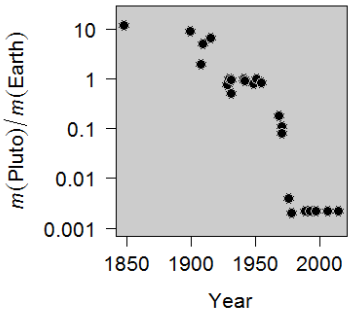
Already in 1848, well in advance of Pluto's discovery, Jacques Babinet estimated the mass of a foretold new planet as 12 times that of Earth. Percival Lowell's 1915 prediction for "planet X" was 6.6 times Earth's mass. And when Clyde Tombaugh finally discovered it in 1930, the world's newspapers announced "a ninth planet, greater than earth, found." Only a few decades ago Pluto was thought to be several orders of magnitude heavier than we now know it to be. What happened that so drastically changed our estimates of Pluto's mass?

Pluto is so distant that it is difficult to learn much about it from direct observation. Our knowledge of its mass therefore depends heavily on the physical models we adopt. For a long time, Pluto's mass was estimated based on perturbations to the motions of Uranus and Neptune. In 1978, however, a sharp-eyed US astronomer, James Christy, discovered Pluto's first moon. At half the size of Pluto, Charon has a significant effect on Pluto's movements and enabled estimating its mass by application of Kepler's laws of planetary motion.

In the late 1980s, the orbits of Pluto and its largest moon Charon were aligned with the line-of-sight from Earth (an arrangement that occurs once in 120 years) which allowed for accurate mass estimates for the first time.[35] In 2015, NASA's *New Horizons* probe flew near Pluto and was able to answer one of the most basic mysteries about Pluto conclusively, estimating its mass to be $0.0022 m_{\text{Earth}}$.

Scientists tend to overestimate the confidence in their results and

Four images from New Horizons Long Range Reconnaissance Imager were combined with color data from the Ralph instrument to create this global view of Pluto in July 2015 — Wikimedia Commons (NASA, 2015).



Estimated mass of Pluto ($m_{\text{Pluto}}/m_{\text{Earth}}$) over the last two centuries [Duncombe and Seidelmann, 1980] serves as a vivid example of how important measurement models and all assumptions that go into these models are in creating knowledge.

[35] R. P. Binzel. Pluto-Charon mutual events. *Geophysical Research Letters*, 16(11):1205–1208, November 1989. doi:10.1029/gl016i011p01205

the quest for the mass of Pluto is not the only example where our collective scientific judgment has fallen short. Determinations of the atomic weights of tellurium and iodine made in the 19th century did not favor Mendeleev's suggestion that the atomic weight of tellurium should be smaller than that of iodine. It is therefore not surprising that the estimates of these two atomic weights should have changed gradually to conform with Mendeleev's suggestion. Although now we know that the atomic weight of tellurium is greater than that of iodine, it is plausible that Mendeleev's pronouncement played an invisible guiding role in contemporary atomic weight measurements of these two elements. This phenomenon is known as the *expectation bias* and it is a reminder that uncertainty estimates are often influenced by unknown effects that have little to do with the measurement they pertain to.

Mount Everest: view from the south — Wikimedia Commons (shrimp01967, 2012).

Height of Mount Everest

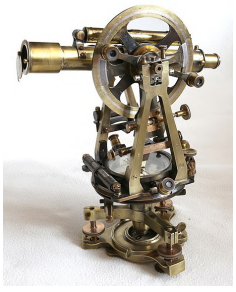
Only in 1849, in the course of the Great Trigonometrical Survey of India (1802–1871), was Mount Everest recognized as the highest mountain on Earth.[36]

The quest to measure the height of Mount Everest reveals how aspects of measurement models that are far too often hidden from view can influence the results. The earliest observations were made from northern India, some 160 km away, and involved measurements of angles made using theodolites. The simplest approach to estimate the height involves only the elevation angle (a), the distance from the observing station to the mountain (d), the altitude of the station (h_S), and a trigonometric relation:

$$h = h_S + d \tan a.$$

For the Jirol station, which stands 67 m above sea level, this formula yields $h = 67\,\mathrm{m} + (190\,966\,\mathrm{m}) \times \tan(1^\circ\,53'\,33.35'') \approx 6377\,\mathrm{m}$, which grossly underestimates the height of the mountain.

[36] S. G. Burrard. Mount Everest: The story of a long controversy. *Nature*, 71:42–46, November 1904. doi:10.1038/071042a0



Troughton & Simms theodolite from around 1910, used to measure angles in horizontal and vertical planes — Wikimedia Commons (Colgill, 2020).

If left uncorrected, the principal sources of error in trigonometrical determinations of height made from long distances are the curvature of the Earth and the refraction of light as it travels through the atmosphere. Accounting for the curvature of the Earth (modeled as a sphere of radius $R = 6371\,\text{km}$) leads to a more complex model:

$$\sin(\tfrac{\pi}{2} - a) = \frac{(R+h)\sin(d/R)}{\sqrt{(R+h_S)^2 - 2(R+h_S)(R+h)\cos(d/R) + (R+h)^2}}.$$

Solving this equation for h numerically, again using the elevation angle measured from the Jirol station, gives $h \approx 9251\,\text{m}$, now overestimating the height.

The fact that atmospheric refraction tends to increase the apparent elevation angle of a mountain peak relative to the observer, is the main reason why the previous height estimate is biased high. While atmospheric refraction depends on several environmental conditions, its magnitude is approximately 10 % of the effect of the curvature of the Earth. The *Manual of Surveying for India* [Thuillier and Smyth, 1875, Page 505] explains how refraction was modeled:

> "There are no fixed rules for Terrestrial refraction, but [...] in determining the heights of the peaks of the Snowy Range (Himalayas), about one-thirteenth of the contained arc was assumed."

Thus, the effect of light refraction was modeled by reducing the observed elevation angle from a to $a - (d/R)/13$ (expressed in radian). As a result, the estimate of the height of Mount Everest, still based on the observation made from Jirol, but now taking into account both the curvature of the Earth and atmospheric refraction, becomes $h \approx 8810\,\text{m}$.

Other influences on the height estimates were recognized later, such as the effect of temperature on the refraction of light and the gravitational influence of these large mountains on plumb lines and leveling devices. Despite all these challenges, the original estimate from the 1850s, 8840 m, is remarkably close to the current estimate of 8848 m, based on GPS measurements made at the mountaintop.

In 1914, *Nature* noted that "when all is said and done, it is the errors arising from the deflection of the plumb-line [...], and the possible variation in the actual height of the point observed (common enough in the case of snow-capped peaks), which chiefly affect the accuracy of angular determinations of altitude, and it is probably to these [...] that we must ascribe [...] the doubt whether Kinchinjunga or K_2 is to hold the honourable position of second in altitude to Everest amongst the world's highest peaks."

STATION	DISTANCE	ANGLE	HEIGHT
Jirol	190.966 km	1° 53′ 33.35″	8836 m
Mirzapur	175.219 km	2° 11′ 16.66″	8841 m
Janjipati	174.392 km	2° 12′ 9.31″	8840 m
Ladnia	175.195 km	2° 11′ 25.52″	8839 m
Harpur	179.479 km	2° 6′ 24.98″	8847 m
Minai	183.081 km	2° 2′ 16.61″	8836 m

Determinations of the height of Mount Everest extracted from the Records of the Great Trigonometrical Survey of India, based on observations made between November 1849 and January 1850 [Burrard, 1904]

The contained arc is the value (in radian) of the angle with vertex at the center of the Earth subtended by an arc of length d on the surface of the Earth. It is the ratio of d to the Earth's radius.

Averaging Models for an Epidemic

In many measurement situations, several alternative models naturally present themselves, with no *a priori* reason to favor one over the others. In some cases it may be most convenient to select and use the "best" model among a collection of alternatives, like we did when we introduced a reliable guide for model building in the context of building a calibration function. In other cases, the best performance is achieved by a weighted average of alternative models.

In general, model averaging does not mean averaging the parameters of the alternative models. The alternative models may have different numbers of parameters, or, even if they have the same number of parameters, the parameters of different models may not be the same kinds of quantities that one could reasonably average. Instead, the averaging will be of predictions that the alternative models make of the same quantities, and the question is how to evaluate the uncertainty of such averages.

The following example illustrates model averaging to produce an estimate of the basic reproduction number (R_0) for an influenza epidemic that ravaged a boarding school for boys between the ages of 10 and 18 in the north of England, during January and February of 1978.

The concept of R_0 is often regarded to be one of the most useful tools in mathematical biology. It is the average number of infections produced by an infective person that interacts freely with others who are susceptible to becoming infected.

DATE	NO. OF CASES
1978-01-22	3
1978-01-23	8
1978-01-24	26
1978-01-25	76
1978-01-26	225
1978-01-27	298
1978-01-28	258
1978-01-29	233
1978-01-30	189
1978-01-31	128
1978-02-01	68
1978-02-02	29
1978-02-03	14
1978-02-04	4

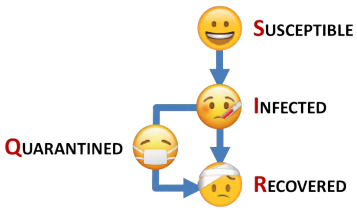
English boarding school epidemic of 1978 [BMJ News and Notes, 1978; Martcheva, 2010]

MEASUREMENT MODELS FOR EPIDEMICS in human or animal populations typically comprise a deterministic component that describes the temporal evolution of the expected number of cases (and the corresponding expected numbers of individuals who are susceptible but not yet sick, of individuals who have already recovered, etc.) [Hethcote, 2000]. These models also comprise a stochastic component that describes how the actual counts of individuals in the different categories vary around their expected values [Bjørnstad, 2018].

The particular epidemic we will be concerned with started in late January and ended in early February of 1978, eventually infecting 512 of the 763 boys in the school. At the peak of the epidemic, 298 boys were confined to bed in the school's infirmary.

We will consider two mathematical models for the daily counts of influenza cases in the boarding school. Their deterministic components are so-called *compartment models*, and their stochastic components are collections of independent Poisson random variables.

At each epoch (a day in this case) a compartment model partitions the relevant population into several categories. For the SIR model these categories are the susceptible, the infective, and the recovered — whose initials, SIR, make the acronym of the model. The SIQR model comprises yet another category, the quarantined. The same

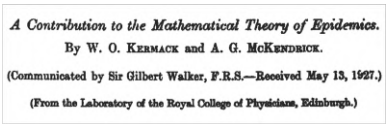


Schematics of two epidemiological, compartmental models of influenza. The SIR model considers only the **SUSCEPTIBLE**, **INFECTED**, and **RECOVERED**, whereas the SIQR model considers also the **QUARANTINED**.

person will belong to different categories at different times as the epidemic spreads and the disease progresses.

We will assume that, at the outset of the epidemic, exactly one boy is infective, and all the others are susceptible. Therefore, the initial counts (on day 1) in the different compartments are

$$S(1) = 762, \quad I(1) = 1, \quad Q(1) = 0, \quad R(1) = 0.$$



A Contribution to the Mathematical Theory of Epidemics.
By W. O. Kermack and A. G. McKendrick.
(Communicated by Sir Gilbert Walker, F.R.S.—Received May 13, 1927.)
(From the Laboratory of the Royal College of Physicians, Edinburgh.)

The SIR model was introduced in the 1920s [Kermack and McKendrick, 1927] and remains one of the simplest models for infectious diseases that are transmitted from human to human, and where recovery confers lasting resistance. This three-compartment model has undergone many improvements and additions tailored for a variety of situations. Recently, for example, the COVID-19 epidemic and the implementation of nationwide interventions in Italy were modeled using an extension of this model that comprises eight compartments: susceptible, infected, diagnosed, ailing, recognized, threatened, healed, and extinct [Giordano et al., 2020]

According to the SIR model, an infected boy will remain infective for some time, and then will recover, in the process acquiring immunity against reinfection with the same virus. But while he is infective, he continues to interact with the other boys in the school, likely spreading the disease.

This is not what actually happened: sick boys were isolated (that is, quarantined) in the school infirmary as soon as the obvious symptoms developed. Quarantining removed them from the pool of those that were spreading the disease. Regardless of whether a sick boy was quarantined or not, eventually he will recover. The SIQR model takes into account the effect of quarantining.

The deterministic components of the SIR and SIQR models are solutions of systems of differential equations, thus assuming that the numbers of boys in the different categories vary continuously over time. The three simultaneous differential equations for the deterministic component of the SIR model are

$$\begin{aligned} dS/dt &= -\beta SI/N, \\ dI/dt &= +\beta SI/N - \gamma I, \\ dR/dt &= +\gamma I. \end{aligned}$$

where $N = 763$ is the total number of boys in the school. Note that S, I and R all are functions of time, t, even if this is not shown explicitly. The observations are the numbers of boys that are sick in bed on each day of the epidemic, which are modeled as outcomes of independent Poisson random variables with means $I(1), \ldots, I(14)$. If the variability of these counts were much in excess of $\sqrt{I(1)}, \ldots, \sqrt{I(14)}$, then a negative binomial model might be preferable.

Since the time derivatives of the numbers of boys in the different compartments add to zero, the total $N = S + I + R$ remains constant over time. More complex models can take into account births and deaths (regardless of whether these are caused by the disease).

The SIQR model has an additional parameter, α, which is the quarantining rate. We assume that the same recovery rate γ applies to all infectives, regardless of whether they are quarantined or not. The SIQR model is represented by the following system of four simultaneous differential equations:

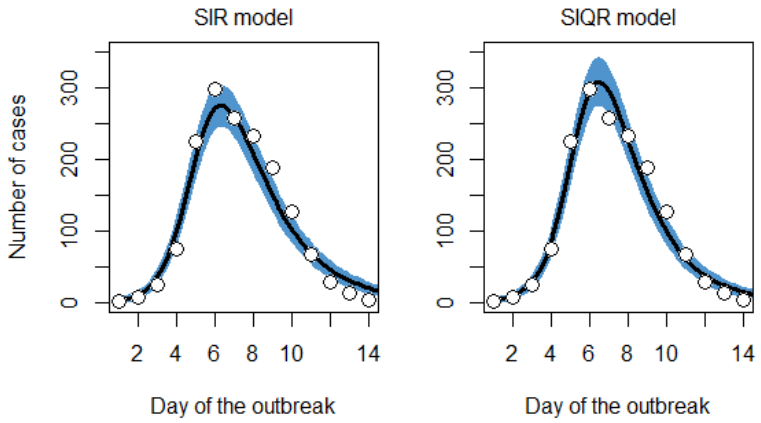
$$\begin{aligned} dS/dt &= -\beta SI/(N - Q), \\ dI/dt &= +\beta SI/(N - Q) - \gamma I - \alpha I, \\ dQ/dt &= +\alpha I - \gamma Q, \\ dR/dt &= +\gamma I + \gamma Q. \end{aligned}$$

These two epidemiological models were fitted to the data using the Stan modeling language, in tandem with the R package `rstan`.[37] The estimates of all non-observable quantities are the means of their Bayesian posterior distributions.

The following R and Stan codes were used to fit the SIR model, assuming that the counts of boys in the different compartments are like outcomes of Poisson random variables whose means satisfy the systems of differential equations presented above.

[37] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017. doi:10.18637/jss.v076.i01; and Stan Development Team. *Stan User's Guide*. `mc-stan.org`, 2019. Stan Version 2.23

```
modelSIR = "functions {
  real[] sir(real t, real[] y, real[] ps, real[] xr, int[] xi) {
    real N = xi[1];
    real dSdt = - ps[1] * y[1] * y[2] / N;
    real dIdt =   ps[1] * y[1] * y[2] / N - ps[2] * y[2];
    real dRdt =   ps[2] * y[2];
    return {dSdt, dIdt, dRdt}; } }
 data { int N; real y0[3]; real ts[14]; int cases[14]; }
 transformed data { real xr[0]; int xi[1] = {N}; }
 parameters { real<lower=0> ps[2]; }
 transformed parameters {
   real y[14,3] = integrate_ode_rk45(sir, y0, 0, ts, ps, xr, xi); }
 model { ps ~ normal(1, 10);  // Priors for beta and gamma
         cases ~ poisson(y[,2]); } // Sampling distribution
 generated quantities { real R0 = ps[1] / ps[2]; }"

library(rstan); library(outbreaks)
cases = influenza_england_1978_school$in_bed
N = 763; n_days = length(cases)
dataSIR = list(n_days=n_days, y0 = c(S=N-1, I=1, R=0),
               N = N, cases = cases, t0 = 0, ts = seq(1, n_days),
               ts_pred = seq(1,1+n_days,length.out = 100) )
## Compile STAN model
modelSIR.poisson = stan_model(model_code=modelSIR)
## Fit STAN model
fitSIR.poisson = sampling(modelSIR.poisson, data = dataSIR)
## Estimate of R0
print(fitSIR.poisson, pars = 'R0')
```



Observed daily numbers of cases and corresponding predicted counts produced by SIR and SIQR models with Poisson variability on the observed cases, surrounded by 95 % probability bands.

The basic reproduction numbers for the SIR and SIQR models are

$$R_0(\text{SIR}) = \frac{\beta}{\gamma},$$
$$R_0(\text{SIQR}) = \frac{\beta}{\gamma + \alpha}.$$

The estimates were $R_0(\text{SIR}) = 3.55$ with $u(R_0(\text{SIR})) = 0.08$ for the SIR model, and $R_0(\text{SIQR}) = 3.38$ with $u(R_0(\text{SIQR})) = 0.08$ for the SIQR model. Although numerically different, they are not significantly different once their associated uncertainties are taken into account: their standardized difference is $(3.55 - 3.38)/\sqrt{0.08^2 + 0.08^2} = 1.5$. Since Bayesian estimates are approximately like outcomes of Gaussian random variables, a z-test for their difference yields p-value 0.13.

The estimates of R_0 produced by these two models can be averaged using Bayesian *stacking weights* [Yao et al., 2017] to produce an estimate corresponding to the best mixture of these models. The weights were computed using R package `loo` [Vehtari et al., 2019]. Since the *stacking weights* were 0.24 for SIR and 0.76 for SIQR, the combined estimate is

$$R_0 = (0.24 \times 3.55) + (0.76 \times 3.38) = 3.42,$$

with uncertainty $u(R_0) = \sqrt{(0.24 \times 0.08)^2 + (0.76 \times 0.08)^2} = 0.06$.

The basic reproduction number, R_0, represents the average number of new infections per existing case. In other words, if $R_0 = 3$, then one person with the disease is expected to infect, on average, three others. Despite its simplicity, R_0 is a *messy* quantity because the definition allows for a multitude of interpretations. For example, do we estimate this quantity at the beginning of the outbreak, at the end, or somehow estimate the average during the entire infectious period?

A common way to estimate R_0, among the many available alternatives,[38] is based on the total number of susceptible patients at the end of the outbreak, which for the boys school was $S(\infty) = 763 - 512 = 251$, using the "final size equation:"

$$R_0 = \frac{\ln(S(0)/S(\infty))}{1 - S(\infty)/N} = \frac{\ln(762/251)}{1 - 251/763} = 1.65.$$

Although *ad hoc*, rather than model-based as the estimates computed above, the very fact that it differs from them to such enormous extent highlights the role of models and the uncertainty that is associated with the selection of a model to estimate the quantities of interest.

R_0 captures various aspects of the outbreak. For simple models such as these, the proportion of the population that needs to be immunized to prevent sustained spread of the disease (that is, to achieve *herd immunity*), has to be larger than $1 - 1/R_0$ and the maximum number of cases on any given day is $I_{\text{MAX}} = N - N(1 + \ln R_0)/R_0$.

For measles, R_0 is widely believed to be somewhere between 12 and 18. Yet, as an example of the real-world *messiness* of the R_0 estimates, a recent systematic review of 18 studies of measles outbreaks reported R_0 values ranging from 4 to 200 [Guerra et al., 2017]

[38] J. M. Heffernan, R. J. Smith, and L. M. Wahl. Perspectives on the basic reproductive ratio. *Journal of The Royal Society Interface*, 2:281–293, 2005. doi:10.1098/rsif.2005.0042

Consensus Building

Burgess and Spangler [2003] explain that "consensus building (also known as collaborative problem solving or collaboration) is a conflict-resolution process used mainly to settle complex, multiparty disputes." In the sciences, consensus building serves to blend measurement results for the same measurand that have been obtained independently of one another. In measurement science in particular, besides this role, consensus building is also used to characterize and compare the different measurement results, by estimating the difference between the true value that each purports to measure, and the true value of the consensus value, and evaluating the corresponding uncertainty — the so-called *degrees of equivalence* [Koepke et al., 2017].

In medicine, where consensus building is often referred to as *meta-analysis* [Higgins et al., 2019], and where the same techniques are also employed to merge results of multicenter trials [Friedman et al., 2015], the goal is to ascertain confidently that a medical procedure or therapy is superior to another, by pooling results from different studies that, if taken individually, may be inconclusive. This approach is also known as *borrowing strength*.

Hubble-Lemaître Constant

In the 1920s, Edwin Hubble and Georges Lemaître discovered that galaxies appear to be moving away from the Earth at speeds (v) that are proportional to their distance (D) from Earth [Hubble, 1929] [Lemaître, 1927, 2013]:

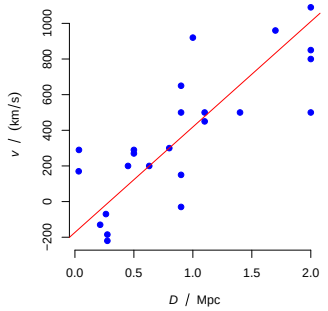
$$v = H_0 D.$$

The constant of proportionality, H_0, is known as the Hubble-Lemaître constant. This discovery motivated Einstein to visit Hubble at the Mount Wilson observatory on January 29, 1931, and acknowledge that the universe is expanding.

The data in Table 1 of Hubble [1929] suggest the estimate $H_0 = 592 \pm 28\,(\mathrm{km/s})/\mathrm{Mpc}$, computed using the procedure described by Bablok et al. [1988] and implemented in R function `deming::pbreg()` (`https://CRAN.R-project.org/package=deming`). As we shall see next, this estimate of H_0 is much larger than contemporary estimates.

Since the final release of the results from the *Planck* survey [Aghanim et al., 2018], which include an estimate of H_0, several other measurement results have been produced for this constant, by application of a wide variety of methods. These results can be combined into a single *consensus* estimate using the following statistical measurement model:

$$H_{0,j} = H_0 + \lambda_j + \varepsilon_j \quad j = 1, \ldots, n$$



Measured values of distance and velocity for 24 galaxies reported by [Hubble, 1929, Table 1]. The red line has slope $592 \pm 28\,(\mathrm{km/s})/\mathrm{Mpc}$.

A parsec is the distance from the Sun to an astronomical object that has a displacement (parallax) angle of one arc second, which is how it got its portmanteau name from parallax and second: it is approximately 3.26 light-years, or 3.1×10^{13} km.

The $\{\lambda_j\}$ denote experiment effects, and the $\{\varepsilon_j\}$ denote measurement errors: the former are modeled as a sample from a Gaussian distribution with mean 0 and standard deviation τ, and the latter are modeled as outcomes of independent Gaussian random variables all with mean 0 and with standard deviations equal to the reported uncertainties $\{u(H_{0,j})\}$.

Neither the $\{\lambda_j\}$ nor the $\{\varepsilon_j\}$ are observable. However, it is possible to tell whether $\tau > 0$ (hence conclude that the $\{\lambda_j\}$ are not all zero) by comparing the dispersion of the measured values $\{H_{0,j}\}$ with the $\{u(H_{0,j})\}$, which are regarded as input data, too. If the measured values are more dispersed than the reported uncertainties suggest they should be, then this means that there are yet unidentified sources of uncertainty in play whose joint effect is accounted for by the introduction of the experiment effects $\{\lambda_j\}$, whose typical size is gauged by τ. Since this "extra" uncertainty manifests itself only when results from multiple, independent experiments are compared, τ is often called *dark uncertainty* [Thompson and Ellison, 2011].

The standard deviation of the thirteen values $\{H_{0,j}\}$ listed alongside (excluding Planck) is 2.8 (km/s)/Mpc, while the median of the $\{u(H_{0,j})\}$, 1.9 (km/s)/Mpc, is 1.5 times smaller. Cochran's Q test of mutual consistency of the measurement results (that is, of the hypothesis that $\tau = 0$) [Cochran, 1954] has p-value 0.016, suggesting that there is significant dark uncertainty.

The model for the $\{H_{0,j}\}$ may be fitted in any one of several different ways. We will use the DerSimonian-Laird procedure as implemented in the the *NIST Consensus Builder* [Koepke et al., 2017], which produces the estimate of dark uncertainty $\widehat{\tau} = 1.92$ (km/s)/Mpc, and consensus value $\widehat{H}_0 = 71.8$ (km/s)/Mpc with $u(\widehat{H}_0) = 0.8$ (km/s)/Mpc.

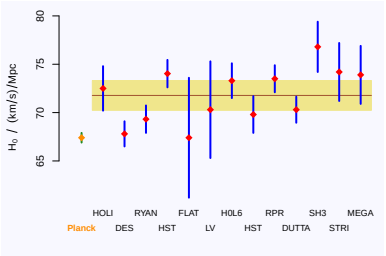
The estimates of H_0 have been steadily shrinking,[39] and by an order of magnitude, from the early values obtained by Hubble and Lemaître, which were in the range 500-600(km/s)/Mpc, to contemporary estimates around the foregoing consensus value.

The *Hubble time*, $t_H = 1/H_0 \approx 14 \times 10^9$ years, may be interpreted as an estimate of the age of the universe, which is believed to lie between t_H and $(2/3)t_H$, depending on the cosmological model. Since H_0 and the age of the universe are related, the value of H_0 may also be inferred from cosmological models. For example, Ryan et al. [2019] treats H_0 as an adjustable parameter when fitting flat and non-flat variants of the ΛCDM model to observations, obtaining estimates for H_0 from 68 (km/s)/Mpc to 75 (km/s)/Mpc.

To compare the cosmological estimate derived from the *Planck* survey with the foregoing consensus value, we compute the normalized

H_0	$u(H_0)$	
67.4	0.5	Planck [Aghanim et al., 2018]
72.5	2.2	H0LI [Birrer et al., 2019]
67.8	1.3	DES [Macaulay et al., 2019]
69.32	1.42	RYAN [Ryan et al., 2019]
74.03	1.42	HST [Riess et al., 2019]
67.4	6.1	FLAT [Domínguez et al., 2019]
70.3	5.15	LV [Hotokezaka et al., 2019]
73.3	1.75	H0L6 [Wong et al., 2019]
69.8	1.9	HST [Freedman et al., 2019]
73.5	1.4	RPR [Reid et al., 2019]
70.3	1.35	DUTTA [Dutta et al., 2019]
76.8	2.6	SH3 [Chen et al., 2019]
74.2	2.85	STRI [Shajib et al., 2019]
73.9	3.0	MEGA [Pesce et al., 2020]

Estimates of the Hubble-Lemaître constant, H_0, all expressed in (km/s)/Mpc. The standard uncertainties are such that each of the intervals $\{H_{0,j} \pm u(H_{0,j})\}$ is believed (by its authors) to include the true value of H_0 with probability 68 % approximately. Some of the uncertainties were originally expressed asymmetrically, but since the asymmetries were very mild, here they have been replaced by the geometric averages of the corresponding, reported "left" and "right" uncertainties.



Each diamond represents a measured value, and each vertical line segment represents an interval $H_{0,j} \pm u(H_{0,j})$. The horizontal line segment represents the consensus value derived from all measurement results except *Planck*'s, and the horizontal, shaded rectangle depicts the associated standard uncertainty.

[39] V. Trimble. H_0: The incredible shrinking constant, 1925-1975. *Publications of the Astronomical Society of the Pacific*, 108:1073–1082, December 1996. doi:10.1086/133837

difference

$$z = \frac{71.8 - 67.4}{\sqrt{0.8^2 + 0.5^2}} = 4.66.$$

On the hypothesis of no difference between the corresponding true values, this normalized difference would be like an outcome of a Gaussian random variable with mean 0 and standard deviation 1. The probability of attaining or exceeding such a difference (regardless of sign) is $p = 3 \times 10^{-6}$, thus suggesting a very significant difference.

This discrepancy, which the astrophysical literature refers to as *Hubble tension* [Poulin et al., 2019], suggests that the pattern of expansion of the universe may have been somewhat more complex than the Hubble-Lemaître "law" contemplates, and indeed may lead to new physics.[40]

[40] J. Sokol. A recharged debate over the speed of the expansion of the universe could lead to new physics. *Science*, March 2017. doi:10.1126/science.aalo877

Arsenic in Kudzu

Kudzu comprises several species of perennial twining vines native to East Asia, which were introduced into the United States in 1876, originally intended as ornamental plants, and subsequently also used as food for cattle and ground cover. Their astonishingly rapid growth and ability to climb and develop roots opportunistically have turned kudzu into a damaging infestation, snuffing other plants large and small, including trees, and covering man-made structures.

Kudzu, "the vine that ate the South." — Kerry Britton, USDA Forest Service, Bugwood.org.

The development of NIST SRM 3268 *Pueraria montana* var. *lobata* (Kudzu) Extract, included an interlaboratory study where 22 laboratories made triplicate determinations of the mass fraction of arsenic in this material, listed and depicted alongside.

The Shapiro-Wilk test of Gaussian shape offers no compelling reason to abandon the hypothesis that all triplets are like samples from Gaussian distributions. Therefore, the triplets will be replaced by their corresponding averages $\{w_j\}$ and associated standard uncertainties $\{u_j\}$ evaluated using the Type A method from the GUM. For example, for laboratory U,

$$w_{\mathrm{U}} = \frac{0.873 + 0.881 + 0.916}{3} = 0.890 \text{ mg/kg},$$

$$u_{\mathrm{U}} = \sqrt{\frac{(0.873 - w_{\mathrm{U}})^2}{3-1} + \frac{(0.881 - w_{\mathrm{U}})^2}{3-1} + \frac{(0.916 - w_{\mathrm{U}})^2}{3-1}} = 0.013.$$

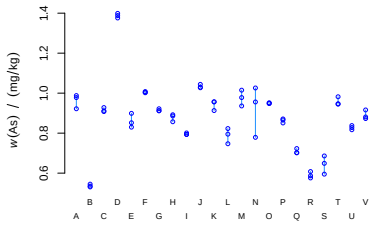
A	0.851	0.866	0.871
B	0.779	0.956	1.026
C	0.702	0.702	0.723
D	0.649	0.686	0.595
E	0.608	0.587	0.576
F	0.899	0.852	0.830
G	0.912	0.912	0.922
H	0.949	0.948	0.952
I	0.947	0.982	0.945
J	0.978	1.015	0.936
K	1.008	1.004	1.002
L	0.908	0.928	0.911
M	1.027	1.030	1.044
N	0.747	0.795	0.823
O	0.801	0.793	0.794
P	0.892	0.886	0.857
Q	0.838	0.817	0.828
R	0.531	0.545	0.535
S	0.922	0.978	0.988
T	1.376	1.399	1.388
U	0.873	0.881	0.916
V	0.913	0.957	0.956

Triplicate determinations of arsenic, where the letters denote laboratories and the numbers are values of mass fraction, expressed in mg/kg.

Cochran's Q-test [Cochran, 1954] suggests significant heterogeneity, even if the determinations made by laboratories B, D, Q, R, and S were to be left out — they will not be left out in our subsequent analyses because there is no substantive reason to.

The symmetry test proposed by Miao et al. [2006] and implemented in R package `symmetry` [Ivanović et al., 2020], applied to the

averages of the triplicates obtained by the participating laboratories, yields p-value 0.37, hence no reason to dismiss a symmetrical model for the random effects. And the Anderson-Darling test of Gaussian shape, applied to the coarsely standardized laboratory-specific averages, yields p-value 0.004. The "coarsely standardized" averages are $\{(w_j - M)/u(w_j)\}$, where M denotes the median of the $\{(w_j\}$, and each w_j is the average of the three replicates obtained by laboratory j, for $j = \mathrm{A}, \ldots, \mathrm{V}$.



Each open circle represents a measured value, and each vertical line segment links the replicates from one laboratory.

Thus, we are faced with a situation where the laboratory-specific lack of repeatability may be modeled using Gaussian distributions, but the laboratory effects require a model that is symmetrical and has tails heavier than Gaussian tails. Consider a random effects model of the form $w_j = \omega + \lambda_j + \varepsilon_j$, where ω denotes the true value of the mass fraction of arsenic in the material, the $\{\lambda_j\}$ have a Laplace distribution with mean 0 and standard deviation τ, and the $\{\varepsilon_j\}$ are Gaussian, all with mean 0 but possibly different standard deviations $\{\sigma_j\}$ that also need to be estimated. The Laplace random effects will dampen the influence that the results from laboratories `B` and `D` will have upon the consensus value and associated uncertainty.

Since the $\{u_j\}$ are estimates of the $\{\sigma_j\}$ based on only 2 degrees of freedom each, and an estimate of τ will be based on the dispersion of only 22 observations, we employ a Bayesian formulation that is best capable to recognize such limitations, and take into account their impact on the evaluation of uncertainty for the consensus value. To this end, we will use the following prior distributions: a largely non-informative, Gaussian prior distribution for ω, a half-Cauchy prior distribution for τ, with median γ, and a half-Cauchy prior distribution for the $\{\sigma_j\}$, with median δ.

Since γ and δ are parameters of prior distributions, they are often called *hyperparameters*. Similarly to how the *NIST Consensus Builder* assigns default values to these parameters, we set γ equal to the `mad` of the laboratory-specific averages, and δ equal to the median of the $\{u_j\}$.

The following Stan and R codes implement and fit the model described above, assuming that the Stan code is in a file called `LaplaceGaussian.stan` located in R's working directory, and that `w` and `u` are vectors of laboratory averages and associated standard uncertainties, and `nu` is the corresponding vector of numbers of degrees of freedom (whose 22 elements all should be equal to 2).

```
library(rstan)
As.Data = list(N=n, w=w, u=u, nu=nu, gamma=mad(w), delta=median(u))
As.Fit = stan(file="LaplaceGaussian.stan", data=As.Data,
              warmup=75000, iter=500000,
              chains=4, cores=4, thin=25)
```

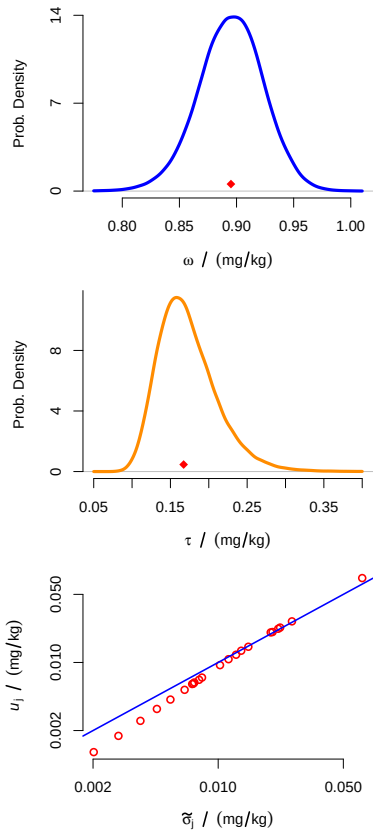
```
data {
  int < lower = 1 > N; // Number of labs
  real gamma;          // Prior median of dark uncertainty tau
  real delta;          // Prior median of {sigma[j]}
  real w[N];           // Measured values
  real u[N];           // Standard uncertainties
  real nu[N]; }        // Numbers of degrees of freedom

transformed data{
  real u2[N];
  for (j in 1:N) {u2[j] = u[j]^2;} }

parameters {
  real omega; real < lower = 0 > tau;
  real theta[N]; real < lower = 0 > sigma[N]; }

model {
  omega ~ normal(0, 100000); // Prior for omega
  // Half-Cauchy prior for tau with median gamma
  tau ~ cauchy(0, gamma);
  // Random effects {theta[j]-omega}
  // Division by sqrt(2) makes tau the prior SD
  theta ~ double_exponential(omega, tau/sqrt(2));
  // Half-Cauchy prior for sigma[j] with median delta
  for (j in 1:N) sigma[j] ~ cauchy(0, delta);
  // Likelihood for u2[j]
  for (j in 1:N) {u2[j] ~ gamma(nu[j]/2,
                                nu[j]/(2*(sigma[j]^2)));}
  // Likelihood for w[j]
  for (j in 1:N) w[j] ~ normal(theta[j], sigma[j]); }
  // Must end with empty line
```

The Stan code treats both the measured values $\{w_j\}$ and the associated uncertainties $\{u_j\}$ as data. Therefore, the likelihood includes a term for the $\{u_j^2\}$ that recognizes the fact that, under the Gaussian assumption for the measured values, the $\{\nu_j u_j^2/\sigma_j^2\}$ are like outcomes of independent random variables with chi-square distributions, which are particular gamma distributions.

An estimate of the posterior probability density of the consensus value ω is depicted in the top panel, alongside. The posterior mean is $\widetilde{\omega} = 0.895\,\mathrm{mg/kg}$, whose associated uncertainty is $u(\widetilde{\omega}) = 0.029\,\mathrm{mg/kg}$. A 95 % credible interval for ω ranges from 0.839 mg/kg to 0.951 mg/kg.

The posterior median for the dark uncertainty, $\widetilde{\tau} = 0.17\,\mathrm{mg/kg}$ is about 20 times larger than the median of the reported uncertainties. The corresponding probability density is depicted in the middle panel.

The bottom panel depicts the relationship between the reported standard uncertainties $\{u_j\}$ and the posterior medians of the corresponding $\{\sigma_j\}$, showing the shrinkage effect that is typical of Bayesian estimates: the posterior median uncertainties tend to be larger than the reported uncertainties for the smaller uncertainties, and smaller than them for the larger uncertainties. In this case, each reported uncertainty is based on only 2 degrees of freedom: the Bayesian model recognizes this limitation explicitly, and takes it into account as it estimates the consensus value and evaluates the uncertainty of the associated uncertainty.



TOP PANEL: Posterior probability density of the consensus value, with mean $\widetilde{\omega} = 0.895\,\mathrm{mg/kg}$ (red diamond). MIDDLE PANEL: Posterior probability density of the dark uncertainty, with median $\widetilde{\tau} = 0.17\,\mathrm{mg/kg}$ (red diamond). BOTTOM PANEL: Reported standard uncertainties, $\{u_j\}$, versus posterior medians of the corresponding $\{\sigma_j\}$.

It is worth noting that nearly identical results can be obtained using the following one-liner from brms package without summarizing the triplicate results from the individual laboratories:

```
library(rstan)
library(brms)
brm(formula = bf(w ~ 1 + 1|lab, sigma ~ 0 + lab, quantile = 0.5),
    family = asym_laplace, data = df)
```

Since the model implemented in `brm` and the model specified above using the Stan language differ only in the choice of prior distributions, the fair agreement of the respective results is a welcome outcome of this sensitivity analysis.

Appendix: Uncertainty

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.

This definition acknowledges explicitly that measurement uncertainty is a kind of uncertainty, and intentionally disconnects the meaning of measurement uncertainty from how it may be represented or described.

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 son was born in the month of February"), or a logical necessity (like, "I am certain that 7253 is a prime number").

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.

Measurement uncertainty can be represented most thoroughly by a probability distribution. This representation applies equally well to the measurement of qualitative as of quantitative properties.

For quantitative, scalar measurands, measurement uncertainty may be summarily, albeit incompletely, represented by the standard deviation of the corresponding probability distribution, or by similar indications of dispersion. A set of selected quantiles of this distribution provides a more detailed summarization than the standard uncertainty.

The uncertainty surrounding quantitative, multivariate or functional measurands, can be summarized by covariance matrices or by coverage regions, for example coverage bands for calibration and analysis functions.

For categorical measurands, the dispersion of the probability distribution over the set of possible values for the property of interest may be summarized by its entropy.[41] Alternatively, the uncertainty may be expressed using rates of false positives and false negatives, sensitivity and specificity,[42] or receiver operating characteristic curves.

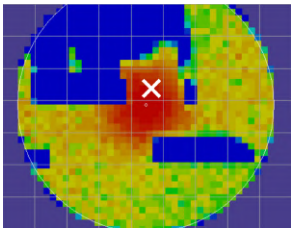
[41] A. Possolo. *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. National Institute of Standards and Technology, Gaithersburg, MD, 2015. doi:10.6028/NIST.TN.1900. NIST Technical Note 1900

[42] D. G. Altman and J. M. Bland. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *British Medical Journal*, 308(6943):1552, 1994. doi:10.1136/bmj.308.6943.1552

When it proves impracticable to express measurement uncertainty quantitatively (either for quantitative or for categorical measurands), it may be expressed using an ordinal scale comprising suitably defined degrees of uncertainty, or levels of confidence. For example, using terms like "Virtually certain", "Very likely", etc., in climatology [Mastrandrea et al., 2011]; or "Most Confident", "Very Confident", etc., in the identification of nucleobases in DNA (NIST SRM 2374 DNA Sequence Library for External RNA Controls), or of a biological species (NIST SRM 3246 *Ginkgo biloba*).

Appendix: Probability

Imagine an explorer looking for the wreckage of an airplane resting at the bottom of the sea, with the aid of a map that is colored with a gradation of colors starting from blue, which indicates the lowest probabilities, then progressing through green, yellow, and orange, and finally reaching red, which indicates the highest probability of the wreckage being there. A *probability distribution* is like this map, or like a distribution of mass over the set of possible values for a measurand: where the colors are reddest, or where the mass density is largest, the more likely it is for the true value of the measurand to be there. For a scalar measurand, the "region" is a subset of the real numbers equipped with the appropriate measurement units. For multivariate measurands, the "region" is a subset of a suitable, multidimensional space. For categorical measurands, the "region" is the set of its possible values.



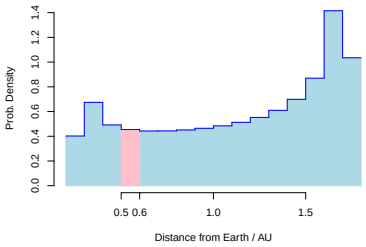
Probability map for the location of Air France 447 site after three unsuccessful searches from June 2009 to May 2010. Red areas indicate highest probability that the wreckage is located there, and the white cross shows the location of where the wreckage was found in 2011. Modified version of Figure 33 in Stone et al. [2011].

Probability distributions over sets of values of quantities or qualities are mathematical objects very similar to distributions of mass in space. Probability, the same as mass, may be distributed continuously, smoothly (as one spreads jelly on bread), or it may be distributed discretely, in lumps (as one places blobs of cookie dough on a baking sheet).

A distribution of probability, like a distribution of mass, may have both features: being smooth in some regions, and lumpy in others. For example, an estimate of dark uncertainty (discussed under *Consensus Building*) typically can be zero with positive probability (hence its probability distribution places a lump of probability at 0), but is otherwise distributed continuously over all positive numbers. The set to which a probability distribution assigns its whole unit of probability is called the *support* of the distribution.

But while different masses may be distributed over the same or different regions of space, all distributions of probability have available a single unit of probability to spread around. Where probability piles up and stands out it suggests where it is more likely that the treasure lies.

Consider the probability density of the distance from Earth to Venus as both planets travel in their orbits around the sun as described by Kepler's laws. The function whose graph is the dark blue polygonal line that tops the histogram is a *probability density function*: it represents probabilities by areas under its graph. The total area painted light blue or pink is 1. The assignment of the unit of probability to the horizontal axis according to the areas under the polygonal line defines what is called a *probability distribution* on this



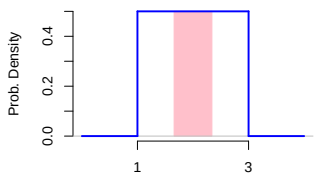
Histogram depicting the probability of finding Venus within particular distances from Earth. The pink rectangle from 0.5 AU to 0.6 AU, has an area 0.04544 which is the probability that, on a randomly chosen day, Venus will be between 0.5 AU and 0.6 AU from Earth. This probability was computed by determining the number of days, between December 25th, 2020, and December 24th, 2420, when the distance to Venus will be in that interval, and dividing it by the total number of days in this period: $6638/146097 = 0.04544$.

axis. In this case, probability piles up toward the ends of the range of distances, and it is scarcer in the middle.

If the area under the polygonal line is conceived as representing matter of uniform density, and this matter collapses to form a rigid rod of negligible thickness on the horizontal axis, then the probability distribution is the distribution of mass of this rod, and the probability density function depicts the variation of the mass density along the rod.

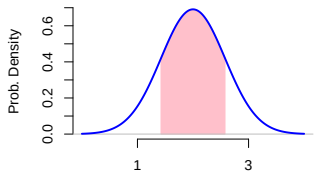
The mean of the distribution is the center of mass of this rod, and the variance of the distribution is the second moment of inertia of the rod when it rotates about its center of mass, with axis of rotation perpendicular to the rod.

Probability distributions naturally arrange themselves into families: Gaussian distributions, Weibull distributions, etc. The members of the same family have probability densities of the same form, differing only in the values of some *parameters*, which identify the individual members of the family. For example, individual Gaussian distributions are identified by the mean and standard deviation, and individual Weibull distributions by the shape and scale parameters.



Probability density of the uniform distribution on the interval $[1, 3]$, with mean 2 and standard deviation $1/\sqrt{3} = 0.58$. The pink region comprises 68 % of the area under the curve.

A uniform (or rectangular) probability distribution over an interval $[a, b]$, where $a < b$, has a probability density function whose value is $1/(b - a)$ over that interval, and zero everywhere else. Since its graph is a rectangle, the distribution is also called *rectangular*. It has mean $\mu = (a + b)/2$ and standard deviation $\sigma = (b - a)/\sqrt{12}$. Since probabilities are given by areas under the graph of the probability density, the probability assigned to an interval $[x - \delta, x + \delta]$, for some $\delta > 0$ and any real number x, either is zero or decreases to zero as δ decreases to zero. For this reason, the distribution is said to be *continuous*. Every continuous distribution thus has the apparently paradoxical property that even though it assigns probability zero to every individual real number, the probability it assigns to its support still adds up to 1.



Probability density of the Gaussian distribution with mean 2 and standard deviation $1/\sqrt{3} = 0.58$. The pink region comprises 68 % of the area under the curve.

A Gaussian (or normal) probability distribution with mean μ and standard deviation $\sigma > 0$ is a continuous distribution whose support is the infinitely long interval that comprises all real numbers. Its probability density has the familiar bell-shaped curve as its graph: it is symmetrical around μ and has inflection points at $\mu \pm \sigma$. The area under the curve between the inflection points is 68 %, and the corresponding area between $\mu \pm 2\sigma$ is 95 % approximately.

The Gaussian distribution plays a central role in probability theory because the probability distribution of the sum of several independent random variables can, under very general conditions, be

approximated by a Gaussian distribution — a remarkable fact first established in fair generality by Pierre Simon, Marquis de Laplace, in 1812.

A unique, surprising property of the Gaussian distribution is that "a necessary and sufficient condition for the normality of the parent distribution is that the sampling distributions of the mean and of the variance be independent."[43] This is surprising because both the sample average and the sample variance are functions of the same data.

[43] E. Lukacs. A characterization of the normal distribution. *Annals of Mathematical Statistics*, 13(1):91–93, March 1942. doi:10.1214/aoms/1177731647

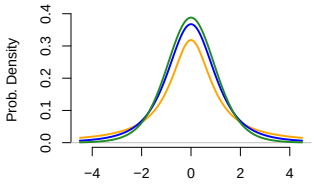
The distribution takes its name from Carl Friedrich Gauss (1777–1855) because he proved that the arithmetic average is the best combination of observations (in the sense of minimizing mean squared error) when the errors of observation are Gaussian, thus providing a rationale for the widespread practice of averaging observations.

"The reference standard for shapes of distribution has long been the shape associated with the name of Gauss, who combined mathematical genius with great experience with the highest-quality data of his day — that of surveying and astronomy. Later writers have made the mistake of thinking that the Gaussian (sometimes misleadingly called normal) distribution was a physical law to which data must adhere — rather than a reference standard against which its discrepancies are to be made plain."
— John Tukey (1977, §19B)

The distribution is also called "normal." However, John Tukey in particular, has made clear that it is far from being a universally adequate model for data. On the contrary, he places the Gaussian distribution among the defining elements of what he calls the *utopian* situation for data analysis — an "ideal" situation that is as mathematically convenient as it often is disjointed from reality.

A Student's t probability distribution with center θ, scale $\tau > 0$, and number of degrees of freedom $\nu > 0$ is a continuous distribution whose support is the set of all real numbers.

The graph of its probability density also is bell shaped, but its tails are heavier (and its center lighter) than in the Gaussian distribution with the same mean and standard deviation. The parameter ν controls its tail heaviness: the smaller the ν, the heavier the tails. For example, Student t-distribution with mean 0 and standard deviation $\sqrt{3}$ (which has 3 degrees of freedom), assigns almost seven times more probability to the interval $[6, 7]$ than a Gaussian distribution with the same mean and standard deviation.



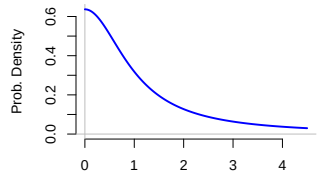
Probability densities of Student's t distributions with center 0, scale 1, and number of degrees of freedom 1 (orange), 3 (blue), and 9 (green).

This distribution is remarkable, and pervasive, owing to this fact: if $x_1, \ldots, x_m$ are a sample of size m drawn from a Gaussian distribution with mean μ and standard deviation σ, $\overline{x}_m$ is the sample average and $s_m^2 = \left((x_1 - \overline{x}_m)^2 + \cdots + (x_m - \overline{x}_m)^2\right)/(m-1)$ is the sample variance, then $(\overline{x}_m - \mu)/(s_m/\sqrt{m})$ is like an outcome of a random variable with a Student's t distribution with center 0, scale 1, and $m-1$ degrees of freedom. This is remarkable because the probability distribution of this ratio does not involve the unknown σ.

If $\nu \leqslant 2$, then the Student's t distribution has infinite variance. A Student's t distribution with $\nu = 1$ is called a Cauchy or Lorentz distribution: it has neither variance nor mean. Random variables with Cauchy distributions are truly wild things. This is how wild:

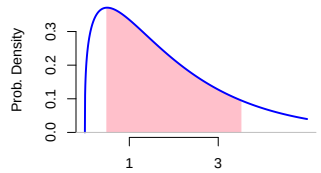
the average of a sample from a Cauchy distribution has the same distribution as the individual sample values.

The Half-Cauchy distribution is the result of truncating at zero a Cauchy distribution that is centered at zero, so that it assigns all of its probability to the positive real numbers. Gelman [2006] suggests the half-Cauchy as a general purpose, weakly informative prior distribution for standard deviations in Bayesian random effects models. We use it in this role when computing a consensus value for the mass fraction of arsenic in kudzu.



Probability density of the Half-Cauchy distribution with median 1.

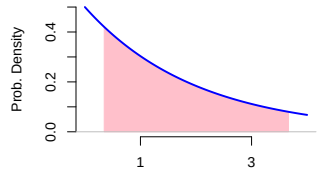
The Gamma and Chi-square probability distributions are members of the same family. The gamma distribution is determined by two positive parameters, shape α and rate β, and its support are the positive real numbers. The distribution is skewed to the right, with a right tail longer and heavier than the left tail. The mean is α/β, and the variance is α/β^2. A gamma distribution with shape $\alpha = 1.7$ and rate $\beta = 762\,\mathrm{kg/mg}$ is used in the measurement of nitrite in seawater to encapsulate prior knowledge about measurement uncertainty associated with Griess's method [Griess, 1879].



Probability density of the gamma distribution with mean 2 and standard deviation $1/\sqrt{3}$. The pink region comprises 68 % of the area under the curve.

If $\nu > 0$, shape $\alpha = \nu/2$, and rate $\beta = ½$, then the gamma distribution is the chi-square distribution with ν degrees of freedom (its sole adjustable parameter), with mean ν and variance 2ν.

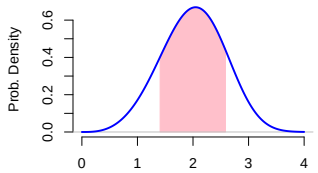
The Gaussian, chi-square, and Student's t distributions are interrelated in a remarkable manner. If $\overline{x}$ and s are the average and standard deviation of a sample of size m drawn from a Gaussian distribution whose mean μ and standard deviation σ both are unknown, then: (i) $\overline{x}$ and s are like outcomes of two independent random variables (even though they are functions of the same data); (ii) $(m-1)s^2/\sigma^2$ is like an outcome of a chi-square random variable with $m-1$ degrees of freedom; and (iii) $(\overline{x}-\mu)/(s/\sqrt{m})$ is like an outcome of a Student's t random variable with $m-1$ degrees of freedom, hence its distribution does not depend on the unknown σ. This last fact is the basis for the coverage intervals specified in Annex G of the GUM [JCGM 100:2008].



The dark blue, steadily decreasing curve is the probability density of the chi-square distribution with both mean and standard deviation equal to 2. The pink region comprises 68 % of the area under the curve. When the number of degrees of freedom ν is greater than 2, the curve has a single hump, reaching a maximum at $\nu - 2$.

The Stan code that implements a random effects model for the determinations of the mass fraction of arsenic in kudzu employs the chi-square distribution in the likelihood function to express the uncertainty associated with sample standard deviations based on small numbers of degrees of freedom as follows: if ν denotes the number of degrees of freedom that s is based on, then $\nu s^2/\sigma^2$ is like an outcome of a chi-square random variable with ν degrees of freedom, and s^2 is like an outcome of a gamma random variable with shape $\nu/2$ and rate $\nu/(2\sigma^2)$.

The Weibull probability distribution may be the most important continuous, univariate distribution, after the Gaussian, chi-square, and Student's t distributions. Its support is the positive real numbers, and it is indexed by two parameters: shape $\beta > 0$ and scale $\eta > 0$, with mean $\eta\Gamma(1+1/\beta)$ and standard deviation $\eta(\Gamma(1+2/\beta) - \Gamma^2(1+1/\beta))^{1/2}$, where "$\Gamma$" denotes the gamma function of mathematical analysis (whose values can be computed in R using function `gamma`). The exponential distribution is a particular case of the Weibull distribution (when the shape is 1). The Weibull distribution is renowned for being an accurate model for the strength of many materials, and for the longevity of mechanical parts and machinery. We will illustrate its use in such setting, when we will describe maximum likelihood estimation and Bayes methods.

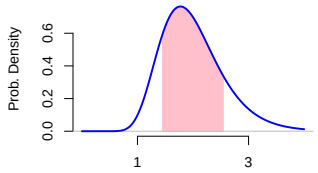


Probability density of the Weibull distribution with mean 2 and standard deviation $1/\sqrt{3} = 0.58$. The pink region comprises 68 % of the area under the curve.

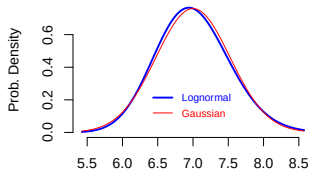
A lognormal probability distribution is a continuous distribution whose support is the positive real numbers. If a random variable X has a lognormal distribution with mean μ and standard deviation $\sigma > 0$, then $\ln(X)$ has a Gaussian distribution with mean $\ln(\mu/\sqrt{(\sigma/\mu)^2+1})$, and variance $\ln((\sigma/\mu)^2+1)$.

Ratios, U/V, arise often in metrology, and the Gaussian distribution just as often is the natural candidate to model the uncertainties that surround them. However, assigning a Gaussian distribution to the denominator, V, implies that the probability is positive that V shall take a value arbitrarily close to zero, hence that the ratio may become arbitrarily large in absolute value, or, in other words, that the uncertainty of the ratio will be infinite. Of course, if zero lies many standard deviations away from V's expected value, then this difficulty may not matter in practice.

When the coefficient of variation of V (standard deviation divided by the mean) is small (say, less than 5 %), then Gaussian and lognormal distributions with identical means and with identical standard deviations will be essentially identical, and the lognormal model may be used to avoid the possibility of inducing an unrealistically large variance for the ratio.

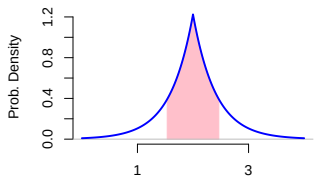


Probability density of the lognormal distribution with mean 2 and standard deviation $1/\sqrt{3} = 0.58$. The pink region comprises 68 % of the area under the curve.



Probability densities of the lognormal (blue thick curve) and Gaussian (red thin curve) distributions, both with mean 7 and standard deviation 0.525. The coefficient of variation is 7.5 %, and the two densities already provide a close approximation to one another.

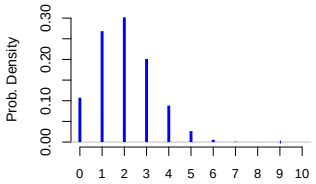
The Laplace probability distribution, also called the double exponential distribution, is specified by its mean and scale parameters. We use the Laplace distribution in a model for the results of an interlaboratory study of the mass fraction of arsenic in kudzu because its tails are heavier than the tails of the Gaussian distribution with the same mean and standard deviation, thus reducing the influence that measured values far from the bulk of the others have upon the consensus value.



Probability density of the Laplace distribution with mean 2 and standard deviation $1/\sqrt{3} = 0.58$. The pink region comprises 68 % of the area under the curve.

A BINOMIAL DISTRIBUTION is often appropriate to characterize the uncertainty surrounding the number of entities of a particular type that are being identified and counted. For example, in a differential leukocyte count where 100 white blood cells were identified and counted, there were 4 eosinophils. This count may be regarded as outcome of a binomial random variable based on 100 *trials* (examinations of individual cells), each of which yields a eosinophil (*success*) or a cell of some other type (*failure*), provided different trials are independent events, and the probability of a cell being identified as an eosinophil remains constant for the duration of the examination.

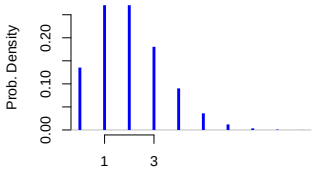
The binomial distribution is indexed by two parameters: the number, n, of "trials", and the probability of "success", $0 \leqslant p \leqslant 1$, in each trial. A random variable with such binomial distribution can take any integer value between 0 and n, inclusive. Its mean, or expected value, is np, and its variance is $np(1-p)$. This variance does not express the component of uncertainty attributable to identification errors: for example, where a true eosinophil is misidentified as being a leukocyte of some other type — the contribution that misidentification makes to combined uncertainty needs to be evaluated separately.



Probability density of the Binomial distribution with mean 2, based on 10 trials, with probability of "success" 0.2 in each trial.

A POISSON PROBABILITY DISTRIBUTION with mean $\lambda > 0$ has standard deviation $\sqrt{\lambda}$. It is a *discrete* distribution because it distributes its unit of probability in lumps at $0, 1, 2, \ldots$. The probability that a Poisson random variable with mean λ will take the value x is $\exp(-\lambda)\lambda^x/x!$, where $x! = x(x-1)(x-2)\ldots 1$. The number of alpha particles emitted per second, as a result of the radioactive disintegration of 1 ng of ^{226}Ra, is a Poisson random variable with mean $\lambda = 36.6$ /s. The numbers of boys that were sick in bed on each day of an influenza epidemic in an English boarding school were modeled as outcomes of independent, Poisson random variables whose means varied from day to day.

The Poisson distribution is often used as a model for the number of occurrences of a rare event because Poisson probabilities can approximate binomial probabilities quite closely, when the probability of "success" is small. A river's 100-year flood is a rare event that occurs once per century on average. This implies that the probability of it occurring on any particular year is 0.01. The binomial probability of it occurring exactly once (meaning once and once only) in a century is $100(0.01)^1(1-0.01)^{99} = 0.3697$. The corresponding Poisson approximation is computed by putting $x = 1$ and $\lambda = 100 \times 0.01 = 1$ in the formula above, to get $\exp(-1)(1)^1/1! = 0.3679$.



Probability density of the Poisson distribution with mean 2.

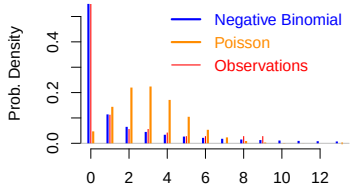
A NEGATIVE BINOMIAL PROBABILITY DISTRIBUTION with mean $\mu > 0$ and dispersion ϕ is a discrete distribution whose support are the non-negative integers. Its variance is $\mu + \mu^2/\phi$. Since the mean and the variance are identical for the Poisson distribution, the presence of the term μ^2/ϕ suggests that the negative binomial may be used as a model for counts whose variance appreciably exceeds their average.

The northern European woodlark (*Lullula arborea*) migrates south in the autumn. These counts were made at the Hanko bird observatory in southwestern Finland, by the Baltic Sea, during the 2009 fall migration season (September-November) [Lindén and Mäntyniemi, 2011], where n_k denotes the number of days with k sightings:

k	0	1	2	3	4	5	6	8	9	17	19	21	25	39
n_k	39	8	4	4	3	2	2	2	2	1	1	1	1	1

For example, in each of 2 days there were 9 sightings of woodlarks at the observatory, and there were 39 days with no sightings. There were, on average, 3.1 sightings per day, and the variance of the number of daily sightings was 44. The negative binomial model, calibrated with maximum likelihood estimates of the parameters, $\widehat{\mu} = 3.1(8)$ and $\widehat{\phi} = 0.22(5)$, fits these data quite well, and incomparably better than the corresponding Poisson model. The overdispersion may be a consequence of the woodlarks' tendency to flock in small groups during autumn.

The woodlark: colored lithograph by Magnus von Wright (1805-1868) — Wikimedia Commons (in the public domain, PD-US-expired)



Probabilities from the negative binomial distribution with mean $\mu = 3.1$ and dispersion $\phi = 0.22$, and from the Poisson distribution with mean $\lambda = \mu$, along with relative frequencies of woodlark sightings.

A MULTINOMIAL PROBABILITY DISTRIBUTION assigns its unit of probability to K different sets or *categories*, so that set $k = 1, \ldots, K$ receives probability $p_k \geqslant 0$ and $p_1 + \cdots + p_K = 1$. Identifying and counting 100 leukocytes is equivalent to placing 100 balls into 7 bins, the balls representing leukocytes and the bins representing the types of leukocytes. The probabilities $\{p_k\}$ may be estimated by the relative frequencies of the different types of leukocytes. In general, if n denotes the number of items to be categorized and counted, then the mean number of items expected for category k is np_k, and the standard deviation of this number is $np_k(1 - p_k)$. The correlation between the numbers of items in categories $1 \leqslant j < k \leqslant K$ is $-\sqrt{p_j p_k/((1-p_j)(1-p_k))}$. Note that all the correlations are negative because an overcount in one category will induce an undercount in another.

RANDOM VARIABLES are quantities or qualities that have a probability distribution as an attribute, and indeed as their most distinctive trait. This attribute serves to indicate which subsets of their respective ranges (the sets where they take their values) are more likely to contain the value that the random variable takes when it is *realized*.

The volume of wine in a bottle of *Volnay Clos des Chênes (Domaine Michel Lafarge)*, from the Côte-d'Or, France, is a random variable that is realized every time a bottle is filled at the winery. The probability distribution of this random variable is continuous, and it is concentrated in a fairly narrow range around 750 mL.

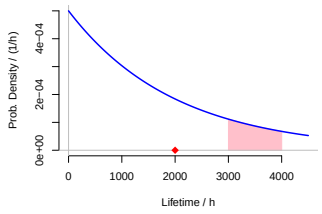
Volnay Clos des Chênes (Domaine Michel Lafarge), from the Côte-d'Or, France

The identity of the nucleotide at a particular locus of a strand of DNA is a random variable whose possible values are adenine, cytosine, guanine, and thymine, and whose realized value is the identity of the nucleotide that is actually there. The probability distribution of this random variable is discrete, its unit of probability being allocated in lumps to those four possible compounds: for example, in the human genome the probability of adenine is 30 %.

The adjective *random* in the expression "random variable X" bears no metaphysical connotation: in particular, it does not suggest that X is an outcome of a game of chance that Nature is playing against us. It is merely a mnemonic and allusive device to remind us that X has a probability distribution as an attribute.

Suppose the random variable in question is the Newtonian constant of gravitation, G, which is generally believed to be constant, but whose true value is known only imperfectly. A probability distribution associated with it can be used to describe the corresponding uncertainty. Similarly, if the quantity is the distance between Venus and Earth, which varies constantly, and in a predictable way, a probability distribution associated with it can describe how it varies over time, or the uncertainty about the distance on a day chosen at random.

The probability distribution of a random variable allows us to compute the probability that it will take a value in any given subset of its range. Doing so is particularly easy when the corresponding probability density is specified analytically. How this is done depends on whether the distribution of the random variable is continuous, discrete, or of a mixed type (that is, has a continuous component over its range, as well as lumps of probability at some of the values in its range).



Probability density of the lifetime of a GE A19 25 W incandescent light bulb. The red diamond marks its expected lifetime, and the area colored pink is the probability that the bulb will last between 3000 h and 4000 h.

Suppose that X is a scalar random variable (for example, the lifetime of a 25 W incandescent light bulb GE A19, whose expected lifetime is 2000 h) and that its probability distribution is continuous and has probability density p_X. Then, the probability that X takes a value in a set A (which may be an interval or a more complicated set), and which we write as $\Pr\{X \in A\}$, is the area under the graph of p_X over the set A. If X has an exponential probability distribution, with density $p_X(x) = \lambda \exp(-\lambda x)$ and mean $\lambda^{-1} = 2000\,\text{h}$ (as depicted alongside), then $\Pr\{3000\,\text{h} < X < 4000\,\text{h}\}$ is the area colored pink,

which in this case can be computed analytically:

$$\int_{3000}^{4000} \frac{1}{2000} \exp(-x/2000) \mathrm{d}x = 0.09.$$

Independence is an important property and a consequential assumption. Two random variables, X and Y, are said to be *independent* when the probability that X takes a value in a set A, and that Y takes a value in a set B, when both are realized jointly, or simultaneously, is equal to the product of their individual probabilities of taking values in such sets one separately from the other.

For example, the number of Aces in a poker hand, and the number of cards from the suit of diamonds in the same hand, are dependent random variables, because knowing that there are five diamonds implies that there cannot be more than one Ace.

Independence is next to impossible to verify empirically in most cases, because doing so involves showing that $\Pr(X \in A \text{ and } Y \in B) = \Pr(X \in A) \times \Pr(Y \in B)$ for all subsets A and B of the respective ranges of X and Y. If these ranges have infinitely many values, then this verification requires an infinitely large experiment.

Two events are independent when the probability of their joint occurrence is equal to the product of their individual probabilities. If the probability of one of them occurring depends on the knowledge of the other one having occurred or not, then the events are dependent.

For example, when rolling two casino dice (perfect cubes with 1, 2, ..., 6 pips on their faces), one red and the other blue, getting 3 pips on the red die, and 7 pips in total, are independent events, but getting 3 on the red die, and 8 (or any other number different from 7) pips in total, are dependent events.

When one says that $\{x_1, \ldots, x_n\}$ is a *sample* from a probability distribution, one means that these are outcomes of n independent, identically distributed random variables whose common distribution is the distribution that the sample allegedly comes from.

The uncertainty of the average of replicated, independent determinations of the same quantity generally will be smaller than the uncertainty of any individual measurement — the prize of claiming independence.
Consider three such determinations with the same standard uncertainty. If modeled as outcomes of independent random variables, then their average will have a standard uncertainty that is $\sqrt{3}$ times smaller than the common standard uncertainty of the individual determinations. If, however, they all are affected by the same error (for example, resulting from miscalibration of the measuring instrument used to obtain them), then averaging the replicates will not reduce the uncertainty.

Exchangeable random variables are such that the random vectors $(X_1, \ldots, X_n)$ and $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ have the same joint probability distribution, for any permutation π of the indices $\{1, \ldots, n\}$. Exchangeable random variables have identical distributions, but generally they are dependent, with correlations never smaller than $-1/(n-1)$.

Exchangeability is often much easier to establish than independence, typically via symmetry arguments. For example, when considering a set of triplicate determinations of the mass fraction of arsenic

The results of five draws (made without replacement) of balls from an urn containing at least five numbered balls, are outcomes of five exchangeable (but dependent) random variables.

in kudzu, (w_1, w_2, w_3), we may conclude that the order in which the determinations were made is irrelevant for any conclusions to be derived from them, hence that they are exchangeable.

Mean, Variance, Bias, and Mean Squared Error are properties of quantitative random variables (or of their probability distributions): the first two (mean and variance) are intrinsic properties of the random variables, and the last two (bias and mean squared error) arise when a random variable plays the role of estimator of a quantity whose true value is unknown.

The mean of a random variable is the center of mass of its probability distribution, when the distribution is regarded as the distribution of a unit of mass over the range of the random variable. And its variance is the second moment of inertia of such distribution of mass, about its mean. The standard deviation is the (positive) square root of the variance: it describes how scattered around the mean the unit of probability is.

The mean is also called the *expected value* (or mathematical expectation), and for this reason the mean of the random variable X is often denoted $\mathbb{E}(X)$. The variance is $\mathbb{V}(X) = \mathbb{E}[X - \mathbb{E}(X)]^2$, and the standard deviation is the (positive) square root of the variance.

If X has a discrete distribution, and the different values that it can take are $x_1, x_2, \ldots$, then $\mathbb{E}(X) = x_1 p_1 + x_2 p_2 + \ldots$, where $p_i = \Pr(X = x_i)$ for $i = 1, 2, \ldots$, provided this sum, which may involve infinitely many summands, is finite.

If X has a continuous distribution, then $\mathbb{E}(X) = \int_{\mathcal{X}} x p(x) \mathrm{d}x$, where p denotes the corresponding probability density, and $\mathcal{X}$ denotes the range of X, provided this integral converges.

Now suppose that a random variable X is to play the role of estimator of a quantity θ whose value is unknown. For example, X may be the mass fraction of inorganic arsenic in a sample of shellfish tissue, and θ may be the true mass fraction of arsenic in it.

Owing to incomplete extraction of the arsenic during sample preparation, the expected value of X may well be less than θ. The *bias* of X as estimator of θ is the difference between its expected and true values, $\mathbb{E}(X) - \theta$. The *mean squared error* (MSE) of X as estimator of θ is the bias squared plus the variance, $(\mathbb{E}(X) - \theta)^2 + \mathbb{V}(X)$.

If X and Y are scalar random variables, and a and b are real numbers, then $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$, regardless of whether X and Y are dependent or independent. And if X and Y are independent, then $\mathbb{V}(aX + bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y)$. In particular, note that $\mathbb{V}(X - Y) = \mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$, provided X and Y are independent. However, nothing like this holds for products, ratios, or any other nonlinear functions of random variables.

Appendix: Statistics

Jimmy Savage [Savage, 1972, Chapter 8] defined "**statistics proper** [. . .] as the art of dealing with vagueness and with interpersonal difference in decision situations." The focus on decision-making suggests an action oriented discipline, "vagueness" refers to uncertainty, whereas the interpersonal difference comprises differences of taste and differences of judgment, both typically varying from person to person.

"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." — GUM 3.4.8 [JCGM 100:2008].

And statistics is an art, similarly to carpentry or cobblery: a practice involving specialized skills and know-how that are developed in apprenticeship with master artisans. Generally not ends in themselves, the statistical arts serve to extract information from data in situations of uncertainty, to enable actions and decisions in all fields of the human endeavor.

Counts

Under *Counting*, we discussed evaluations of uncertainty for counted quantities: numbers of white blood cells (leukocytes) of different types, in particular. We considered a sample of 100 leukocytes comprising 4 eosinophils.

If this count should be modeled as an outcome of a binomial random variable that counts the number of "successes" in 100 independent trials with probability of "success" $4/100$, then the corresponding standard uncertainty will be $\sqrt{100 \times (4/100) \times (96/100)} = 1.96$. The Poisson model that approximates this binomial distribution has mean $100 \times (4/100) = 4$, hence standard deviation 2.

A method proposed by Wilson [1927][44] to build confidence intervals for binomial proportions performs quite well in general.[45] For the true proportion of eosinophils, based on the aforementioned observed count of 4 in a sample of 100, it produces a 95 % confidence interval ranging from 0.013 to 0.11 (thus asymmetrical relative to the observed proportion, 0.04), obtained by executing the R command `prop.test(x=4, n=100)$conf.int`.

[44] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212, 1927. doi:10.2307/2276774

[45] R G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8):857–872, 1998. doi:10.1002/(sici)1097-0258(19980430)17:8<857::aid-sim777>3.0.co;2-e

The uncertainty analysis reported for eosinophils takes two sources of uncertainty into account: sampling variability and between-examiner variability.

Sampling variability is modeled using a multinomial model, to take into account the fact that the counts of the different types of leukocytes are like outcomes of dependent, binomial random variables.

Between-examiner variability is modeled using Gaussian distributions (one for each kind of leukocyte), all with mean zero and with

standard deviations that depend on the type of leukocyte, and are set equal to the standard uncertainties that Fuentes-Arderiu et al. [2007] evaluated. These Gaussian "errors" are added to the counts simulated using the multinomial distribution, using a Monte Carlo method.

Bootstrap

The statistical bootstrap[46] is a computationally-intensive method for statistical inference, and in particular for uncertainty evaluation. Diaconis and Efron [1983] provide a compelling, accessible introduction to the bootstrap, and Hesterberg [2015] describes bootstrapping techniques, copiously illustrated with examples.

There are two main versions of the bootstrap: parametric and non-parametric. Both can be applied to univariate and multivariate data (for example, for the scores in the Ladies Single Skating competition of the 2018 Winter Olympics, and for the calibration of a GC-MS instrument used to measure concentration of chloromethane). Here we begin with a set of replicated determinations $x_1, \ldots, x_m$ of a scalar quantity, obtained under conditions of repeatability.

[46] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, UK, 1993; and A. C. Davison and D. Hinkley. *Bootstrap Methods and their Applications*. Cambridge University Press, Cambridge, UK, 1997. ISBN 0-521-57471-4. URL statwww.epfl.ch/davison/BMA/

The parametric bootstrap regards these determinations as if they were a sample from a probability distribution P_θ that is indexed by a possibly multidimensional parameter θ. The underlying assumption is that this distribution is an adequate model for the variability of the replicates. We also assume that the true value of the measurand, $\eta = \psi(\theta)$, is a known function ψ of θ.

The parametric bootstrap involves three steps:

(PB1) Estimate θ using the observations $\{x_i\}$, to obtain $\widehat{\theta}$.

(PB2) Draw a large number, K, of samples of size m from $P_{\widehat{\theta}}$, and compute the estimate of θ for each of these samples, obtaining $\theta^*_1, \ldots, \theta^*_K$.

(PB3) Compute the corresponding estimates of the measurand, $y_1 = \psi(\theta^*_1), \ldots, y_K = \psi(\theta^*_K)$, and use them as if they were a sample drawn from the distribution of the measurand, to evaluate the associated uncertainty.

Note that we are sampling from $P_{\widehat{\theta}}$: that is, pretending that $\widehat{\theta}$ is θ (which is unknown). K should be no smaller than 10^3 when the method is used to compute standard deviations of functions of the data, and ideally of the order of 10^6 for most purposes.

The standard deviation of the $\{y_k\}$ is an evaluation of standard uncertainty, and the 2.5th and 97.5th percentiles of the $\{y_k\}$ are the endpoints of a coverage interval for the true value of the measurand, with 95 % probability.

The parametric bootstrap is used below to evaluate the uncertainty associated with the maximum likelihood estimate of the tensile strength of alumina coupons in a 3-point flexure test.

The non-parametric bootstrap requires that a "recipe" R be available to combine the replicated observations and produce an

estimate of the measurand: $y = R(x_1, \ldots, x_m)$. This "recipe" may be as simple as computing their median, or it may be an arbitrarily complicated, nonlinear function of the data. The observations again are regarded as a sample from some probability distribution, but here this distribution remains unspecified (hence the qualifier *non-parametric*).

The non-parametric bootstrap is even bolder than the parametric one. For the parametric bootstrap we estimated a parameter of a probability distribution, and proceeded to sample from this distribution pretending that the estimate of the parameter is equal to the true value of the parameter. For the non-parametric bootstrap we will treat the set of replicates in hand as if it were an infinitely large sample from the unspecified, underlying probability distribution, by taking these steps:

(NPB1) Select a large, positive integer K, and for each $k = 1, \ldots, K$ draw $s_{1k}, \ldots, s_{mk}$ uniformly at random, and with replacement, from the set $\{x_1, \ldots, x_m\}$. Each s_{ik} is equal to one of the $\{x_i\}$. For each k, the $\{s_{ik}\}$ are called a *bootstrap sample*.

This means that we get s_{1k} by drawing one of the observations we have as if drawing a ball from a lottery bowl, and then return it back to the bowl, mix the contents, and then draw the observation that will become s_{2k} and so on. Note that the same observation may appear multiple times in a bootstrap sample.

(NPB2) For each bootstrap sample, compute the corresponding estimate of the measurand, $y_k = R(s_{1k}, \ldots, s_{mk})$, and then use the $\{y_k\}$ to evaluate the associated uncertainty, similarly to what was suggested above, under (PB3).

The number K should be as large as practicable, the guidelines being the same as offered above, for the parametric bootstrap. When applying the bootstrap, the first thing to do is to examine the probability distribution of the bootstrap estimates of the measurand, $\{y_k\}$, for example by building a histogram of these values (if the measurand indeed is a scalar quantity).

If this distribution is very "lumpy", with only a few different values, then the bootstrap may not produce a reliable uncertainty evaluation. This may happen when the number m of observations is small, or when the way of combining them tends intrinsically to produce a small number of different values (this can happen, for example, if $R(x_1, \ldots, x_m)$ is the median of the $\{x_i\}$).

In general, m should be large enough for there to be an appreciable number of possible, different bootstrap samples. This can be the case even when m is surprisingly small, because given a set of m observations whose values are all different from one another, it is possible to form $\binom{2m-1}{m-1} \approx 2^{2m-1}/\sqrt{m\pi}$ different bootstrap samples using the non-parametric bootstrap.

For $m = 14$ (the number of replicated determination of the mass fraction of magnesium discussed below), the number of different bootstrap samples is already over 20 million (of course, not all of

these bootstrap samples produce different estimates of the measurand). It is very unlikely that, with $m < 12$, the non-parametric bootstrap will produce reliable results even when the estimate of the measurand is highly sensitive to each single observation. Chernick [2008] suggests that the number of observations should be at least 50.

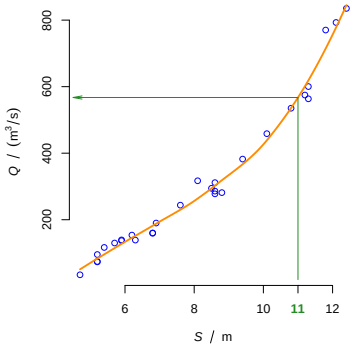
Under *Combining Replicated Observations*, we apply the non-parametric bootstrap to evaluate the uncertainty associated with the median of the Walsh averages (Hodges-Lehmann estimator), using facilities available in R package `boot` [Canty and Ripley, 2020]. Here, we illustrate the non-parametric bootstrap without resorting to these facilities, to make transparently clear what is involved.

River flood *stage* (S) is the height of the water surface above a reference level, and *discharge* (Q) is the volumetric flow rate. The record of yearly peak discharges in the Red River of the North, for the period 1989–2018, and the corresponding flood stages measured at Fargo, North Dakota, can be used to calibrate a relationship between flood stage and discharge, so that flood stage, which is easier to measure accurately than discharge, can be used to estimate discharge.

Year	S	Q	Year	S	Q
1989	10.8	535.2	2005	8.6	277.8
1990	4.7	34.5	2006	11.3	563.5
1991	5.2	74.5	2007	9.4	382.3
1992	5.2	73.3	2008	5.9	137.1
1993	8.6	286.0	2009	12.4	835.3
1994	8.1	317.1	2010	11.3	600.3
1995	8.6	311.5	2011	11.8	770.2
1996	8.8	281.5	2012	5.4	116.7
1997	12.1	792.9	2013	10.1	458.7
1998	7.6	243.8	2014	8.5	294.5
1999	6.3	138.8	2015	5.9	139.3
2000	6.8	159.4	2016	5.2	95.4
2001	11.2	574.8	2017	6.8	160.3
2003	6.9	190.0	2018	5.7	130.0
2004	6.2	153.8			

Values of flood stage (S), expressed in meters above the reference level, and discharge (Q), expressed in m^3/s, at the yearly peak discharge, for the Red River of the North, measured at Fargo, North Dakota (USGS station 05054000).

```
S = c(10.8, 4.7, 5.2, 5.2, 8.6, 8.1, 8.6, 8.8, 12.1, 7.6,
      6.3, 6.8, 11.2, 6.9, 6.2, 8.6, 11.3, 9.4, 5.9, 12.4,
      11.3, 11.8, 5.4, 10.1, 8.5, 5.9, 5.2, 6.8, 5.7)
Q = c(535.2, 34.5, 74.5, 73.3, 286, 317.1, 311.5, 281.5,
      792.9, 243.8, 138.8, 159.4, 574.8, 190, 153.8, 277.8,
      563.5, 382.3, 137.1, 835.3, 600.3, 770.2, 116.7, 458.7,
      294.5, 139.3, 95.4, 160.3, 130)
z = data.frame(S=S, Q=Q)
```



Relation between discharge (Q) and flood stage (S) for the Red River of the North, at the yearly peak discharge, for the period 1989-2018.

The following R code fits a non-parametric, locally quadratic regression model, *loess*,[47] which expresses discharge as a function of flood stage, and then uses the fitted model to estimate the discharge that corresponds to flood stage $S = 11\,\mathrm{m}$: $\widehat{Q}(11\,\mathrm{m}) = 567.4\,\mathrm{m}^3/\mathrm{s}$. The R function `predict` evaluates the associated standard uncertainty as $u(\widehat{Q}(11\,\mathrm{m})) = 9.3\,\mathrm{m}^3/\mathrm{s}$.

```
z.loess = loess(Q~S, data=z)
Q11.loess = predict(z.loess, newdata=data.frame(S=11), se=TRUE)
```

The non-parametric bootstrap, implemented below, involved drawing 10 000 samples, each of size 29, from the set of 29 pairs of observations $\{(S_i, Q_i)\}$, with replacement, fitting the *loess* model to each such sample, and then using the fitted model to predict the discharge corresponding to $S = 11\,\mathrm{m}$. The standard deviation of the resulting 10 000 predicted values of the discharge, $13.1\,\mathrm{m}^3/\mathrm{s}$, is an alternative, 41 % larger and more realistic evaluation of $u(\widehat{Q}(11\,\mathrm{m}))$ than the evaluation derived from the original *loess* fit.

[47] W. S. Cleveland, E. Grosse, and W. M. Shyu. Local regression models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, chapter 8. Wadsworth & Brooks/Cole, Pacific Grove, California, 1992

```
Q11.boot = numeric(10000)
for (k in 1:10000) {
    iB = sample(1:29, size=29, replace=TRUE)
    zB.loess = loess(Q~S, data=z, subset=iB)
    Q11.boot[k] = predict(zB.loess, newdata=data.frame(S=11)) }
c(mean(Q11.boot, na.rm=TRUE), sd(Q11.boot, na.rm=TRUE))
```

Combining Replicated Observations

Consider the problem of estimating the mass fraction of magnesium in a breakfast cereal, based on 14 determinations made using inductively coupled plasma optical emission spectroscopy (ICP-OES), under conditions of repeatability, which are expressed in mg/kg — 1130.0, 1083.3, 1091.7, 1072.0, 1083.2, 1014.6, 1068.0, 1125.6, 1124.6, 1115.3, 1088.1, 1075.0, 1126.8, 1121.1. (These, together with other measurement results, were used to produce the certified value of the mass fraction of magnesium in NIST SRM 3233.)

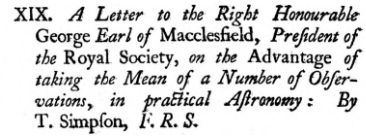
Choosing to minimize the mean squared difference between the estimate and the true value, or to minimize the absolute value of this difference, are different options that can be interpreted as means to achieve optimal estimation under different assumptions: that these determinations are either a sample from a Gaussian distribution, or a sample from a Laplace distribution. The former suggests the arithmetic mean, the latter the median. However, many other modeling choices are conceivable, each leading to a different estimate.

"The problem of summarizing the location of a single batch of numbers is surely the simplest and most classical of the problems recognized as analysis of data. It was first attacked about 1580, by the use of the arithmetic mean. The next few centuries included the statement and proof of the Gauss-Markoff theorem which asserted the minimum-variance property — *among all unbiased estimates linear in the data* — in any problem where the parameters entered linearly into the average value of each observation, for the results of linear least squares. Since the use of an arithmetic mean to summarize a batch was a special instance of this general theorem, the naive might conclude that the problem of summarizing a batch had been settled. Far from it."

— John Tukey (1986)

THE SIMPLE AVERAGE, or arithmetic mean, is the optimal estimate if one chooses to gauge performance in terms of mean squared error, and if one judges the following model to be adequate for the observations: $w_i = \omega + \varepsilon_i$ for $i = 1, \ldots, m$, where $m = 14$ is number of observations, ω is the true value of that mass fraction, and the $\{\varepsilon_i\}$ are measurement errors regarded as a sample from a Gaussian distribution with mean 0 and standard deviation σ. The statistical model, as just formulated, involves the assumption that the observations are not persistently offset from the true value they aim to estimate. This is formalized in their mathematical expectation being equal to the true value: $\mathbb{E}(W_i) = \mathbb{E}(\omega) + \mathbb{E}(\varepsilon_i) = \omega$ because ω is a constant, and the assumption was made above that $\mathbb{E}(\varepsilon_i) = 0\,\mathrm{mg/kg}$. Note that here we have used W_i, the uppercase version of w_i, to denote the random variable that the observation w_i is regarded as a realized value of. Since the expected value of each W_i is ω, we say that there is no *bias* (persistent, or systematic error) in the measurement.

The assumption that the measurement errors $\{\varepsilon_i\}$ are Gaussian implies that so are the $\{w_i\}$, which can be tested. The Shapiro-Wilk[48] and the Anderson-Darling[49] tests, for conformity of a sample with



XIX. *A Letter to the Right Honourable* George *Earl of* Macclesfield, *Prefident of the* Royal Society, *on the* Advantage *of taking the Mean of a Number of Obfervations, in practical Aftronomy: By* T. Simpfon, *F. R. S.*

In his 1755 letter to the Royal Society, Thomas Simpson, Professor of Mathematics at the Royal Academy at Woolwich, outlined the advantages of averaging observations. As an example, the probability that the mean of six observations will have a larger absolute error than a single observation is only 25 % when the errors follow Gaussian distribution. [Simpson, 1755] (CREDIT: archive.org).

[48] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3,4): 591–611, 1965. doi:10.2307/2333709

[49] 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. doi:10.1214/aoms/1177729437

a Gaussian distribution, are commonly used: in this case, the former yields a p-value of 0.06, and the latter of 0.1. It is a common convention in science that only p-values smaller than 0.05 indicate a statistically significant discrepancy, but this is a matter of (subjective) judgment. Indeed, one cannot identify a single universal threshold of statistical significance, and some argue that the level of significance should be set at 0.005.[50]

[50] V. E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110:19313–19317, 2013. doi:10.1073/pnas.1313476110

Suppose that, to test a hypothesis H (in a significance test) one rejects H when the value of some test criterion (a suitable function of the data) is too large. The p-value of the test is the probability, computed on the assumption that H is true, of observing a value of the test criterion at least as large as the value that was obtained using the data available for the test. Since a small p-value suggests that the data are unlikely if H is true, the common practice is to reject H in such case. Of course, one needs to decide in advance how small the p-value needs to be to warrant rejecting H.

The median of the observations is responsive to choices different from those that suggest the average. That instead of seeking to minimize mean squared error, one wishes to minimize mean absolute error, which may be particularly appropriate when the measurement errors $\{\varepsilon_i\}$ have a probability distribution with heavier tails than the Gaussian: for example, Laplace (also known as double exponential). The median is found by ordering the observations from smallest to largest, and selecting the middlemost (when the number of observations is odd), or the average of the two middlemost ones (when the number of observations is even).

For the determinations listed above, the average is 1094.2 mg/kg, and the median is 1089.9 mg/kg. The average has one serious shortcoming: it offers no protection against the influence of a single value that, for one reason or another, lies far from the bulk of the others. Suppose that, owing to a clerical error, the last value is reported as 11 211 mg/kg instead of 1121.1 mg/kg: in consequence, the average will shoot up to 1814.9 mg/kg, while the median stays put at 1089.9 mg/kg.

But the median is also open to criticism. First, it seems to gloss over most of the information in the data: it uses the data only to the extent needed to determine which is the middlemost value. Second, it is sensitive to small perturbations of the middlemost observations. Suppose that the last two digits of the third determination, 1091.7 mg/kg, are transposed accidentally, and 1097.1 mg/kg is reported instead. The average hardly budges, becoming 1094.6 mg/kg, while the median slides to 1092.6 mg/kg.

[51] J. L. Hodges and E. L. Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34(2):598–611, June 1963. doi:10.1214/aoms/1177704172

The median of the Walsh averages (better known as the Hodges-Lehmann estimate[51]) affords a fairly general, flexible solution to the problem of combining replicated observations. It is computed by taking these three steps for a sample of size m:

(1) Compute the averages of all different subsets with two observations each (since two subsets are identical if they have the same elements regardless to order, there are $\frac{1}{2}m(m-1)$ such subsets);

(2) Form a set with these averages together with the m observations;

(3) Find the median of the $\frac{1}{2}m(m+1)$ values in this set.

The Walsh averages are these: $\{(w_i + w_j)/2 : 1 \leqslant i \leqslant j \leqslant m\}$, thus including averages like $(1130.0 + 1130.0)/2$ and $(1130.0 + 1083.3)/2$, but not both $(1083.3 + 1130.0)/2$ and $(1130.0 + 1083.3)/2$, because $\{1130.0, 1083.3\}$ and $\{1083.3, 1130.0\}$ are the same subset.

The following facts make the Hodges-Lehmann estimate particularly attractive, and an excellent, general purpose replacement for the average and the median, particularly when the replicated observations may be assumed to be a sample from a symmetrical distribution:

- It uses the information in the data almost efficiently as the average, when the average is at its best;
- It can use the information in the data far more efficiently than the average, when the average is not at its best;
- It is resistant to outliers;
- It is easy to compute its standard uncertainty, as well as expanded uncertainties and coverage intervals for different coverage probabilities, for example using R:

```
w = c(1130.0, 1083.3, 1091.7, 1072.0, 1083.2, 1014.6, 1068.0,
      1125.6, 1124.6, 1115.3, 1088.1, 1075.0, 1126.8, 1121.1)
w68 = wilcox.test(w, conf.int=TRUE, conf.level=0.68)
HL = w68$estimate; names(HL) = NULL
uHL = diff(w68$conf.int)/2
w95 = wilcox.test(w, conf.int=TRUE, conf.level=0.95)
U95HL = diff(w95$conf.int)/2
Lwr95 = w95$conf.int[1]; Upr95 = w95$conf.int[2]
c(HL=HL, "u(HL)"=uHL, "U95(HL)"=U95HL, Lwr95=Lwr95, Upr95=Upr95)
```

For the 14 replicates of the mass fraction of magnesium, the median of the Walsh averages is 1098.8 mg/kg, with standard uncertainty 9.4 mg/kg, and expanded uncertainty for 95 % coverage 18 mg/kg. Their counterparts for the average are 1094.2 mg/kg, 8.6 mg/kg, and 19 mg/kg, respectively.

And for the median, using the non-parametric statistical bootstrap as implemented in the following R code, gives standard uncertainty 14 mg/kg and expanded uncertainty 24 mg/kg:

```
miB = replicate(1e5, median(sample(w, 14, replace = TRUE)))
u = sd(miB)
U95 = diff(quantile(miB, c(0.025,0.975)))/2
c("u(median)"=u, "U95(median)"=U95)
```

Weighted averages may be appropriate under the same general conditions that make the average optimal, but when the different observations being combined have different uncertainties, for example in the case of the determinations of equivalent activity reported for ^{59}Fe in a key comparison organized by the BIPM using the International Reference System.[52] The synthetic radionuclide ^{59}Fe has half-life of 44.5 days, and decays to stable ^{59}Co via beta decay.

The weighted average of values $x_1, \ldots, x_m$, with non-negative weights $w_1, \ldots, w_m$ (which do not necessarily sum to 1), is

$$\overline{x}_w = \frac{x_1 w_1 + \cdots + x_m w_m}{w_1 + \cdots + w_m}.$$

[52] C. Michotte, G. Ratel, S. Courte, K. Kossert, O. Nähle, R. Dersch, T. Branger, C. Bobin, A. Yunoki, and Y. Sato. BIPM comparison BIPM.RI(II)-K1.Fe-59 of activity measurements of the radionuclide 59Fe for the PTB (Germany), LNE-LNHB (France) and the NMIJ (Japan), and the linked APMP.RI(II)-K2.Fe-59 comparison. *Metrologia*, 57(1A):06003, January 2020. doi:10.1088/0026-1394/57/1a/06003

If $w_i = 1/u^2(x_i)$ then

$$u_{\mathrm{I}}(\overline{x}_w) = \frac{1}{\sqrt{1/u^2(x_1) + \cdots + 1/u^2(x_m)}}.$$

This is the so-called "internal" estimate of $u(\overline{x}_w)$. The "external" estimate is based on the weighted standard deviation of the observations:

$$u_{\mathrm{E}}(\overline{x}_w) = \sqrt{\frac{w_1(x_1 - \overline{x}_w)^2 + \cdots + w_m(x_m - \overline{x}_w)^2}{m(w_1 + \cdots + w_m)}}.$$

The weighted average of the measured values of the equivalent activity of radionuclide ^{59}Fe is 14 619 kBq. The "internal" standard uncertainty is 10 kBq, and the "external" standard uncertainty is 19 kBq. The non-parametric statistical bootstrap, applied to the weighted average of these equivalent activities, produces standard uncertainty 21 kBq, suggesting that $u_{\mathrm{E}}(\overline{x}_w)$ is the more realistic assessment, even if biased low.

The "internal" and "external" evaluations of the standard uncertainty are very different in this case because the measurement results are mutually inconsistent, exhibiting substantial dark uncertainty (explained under *Consensus Building*), and should not be combined using either the simple average or the weighted average with weights inversely proportional to the squared reported uncertainties.

LAB	YEAR	ACTIVITY
BKFH	2001	14 685(32) kBq
IAEA/RCC	1978	14 663(24) kBq
PTB	2012	14 609(25) kBq
NIST	2001	14 641(60) kBq
NPL	1979	14 668(55) kBq
ANSTO	1980	14 548(54) kBq
CMI-IIR	1984	14 709(36) kBq
LNE-LNHB	2013	14 603(36) kBq
NMIJ	2014	14 576(23) kBq
BARC	1998	14 511(28) kBq
KRISS	1999	14 728(50) kBq

The measurement results for equivalent activity, A_{e}, of iron-59 from continuous long-term interlaboratory study Michotte et al. [2020].

WEIGHTED MEDIANS are preferable to the simple median when the observations being combined have different uncertainties, and the median is appropriate to begin with. R function `weighted.median`, as defined in package `spatstat`,[53] computes the weighted median correctly. The weighted median of the measured values of the equivalent activity of ^{59}Fe is 14 606 kBq. The associated standard uncertainty and expanded uncertainty for 95 % coverage, computed using the non-parametric statistical bootstrap, are 31 kBq and 54 kBq, respectively.

[53] A. Baddeley and R. Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42, 2005. URL www.jstatsoft.org/v12/i06/

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a technique used to estimate the values of parameters that appear in observation equations (that is, statistical measurement models). MLE may be used to estimate an input quantity that appears in a conventional measurement model as specified in the GUM, based on replicated observations, or it may be used to estimate the output quantity if the measurement model lends itself to such treatment.

MLE produces not only an estimate of the quantity of interest, but it produces also an approximate evaluation of the associated uncertainty. And if supplemented with the statistical bootstrap, when

this is practicable, then it can characterize uncertainty much more accurately than the approximation to standard uncertainty described in the GUM. The idea of supplementing MLE with the bootstrap is illustrated below, in relation with the measurement of the tensile strength of alumina.

In its most succinct and general form, a statistical measurement model comprises these two statements:

$$\begin{aligned} &(1) \quad X \sim P_\theta; \\ &(2) \quad \eta = \varphi(\theta). \end{aligned}$$

In the first statement, $X = (X_1, \ldots, X_n)$ is a vector of random variables whose probability distributions characterize their uncertainties. Statement (1) says that the joint probability distribution of these random variables is P_θ, where the true value of the parameter θ (typically also a vector, but with a number of components that does not vary with n) is an unknown element of a set H. Statement (2) says that η, denoting the true value of the measurand (which may be a vector), is a known function φ of θ.

Now, suppose that P_θ has probability density p_θ, and that x is the observed value of the vector X. The maximum likelihood estimate of θ is $\widehat{\theta}$ that maximizes $p_\theta(x)$ as θ ranges over H: the idea is to choose a value for the parameter θ that makes the data "most likely." The MLE of the measurand is $\widehat{\eta} = \varphi(\widehat{\theta})$.

In this process, x is kept fixed at its observed value, while θ is allowed to vary until a maximum of $p_\theta(x)$ is found. To emphasize this fact, one often defines a function L_x, called the *likelihood function*, as follows: $L_x(\theta) = p_\theta(x)$. None of the pieces changes, only the viewpoint: the subscript x in L_x is a way of saying that L_x depends on x but that x remains fixed while we seek to maximize $L_x(\theta)$ by varying its argument, θ, over the set H of its possible values. In applications, the subscript x is often suppressed because the dependence on x is understood, and one writes simply $L(\theta)$.

Therefore, maximum likelihood estimation amounts to maximizing the likelihood function. In some cases this can be done analytically, based on the first and second derivatives of L_x with respect to θ. In other cases it has to be done via numerical optimization.

Under very general circumstances, maximum likelihood estimation enjoys several remarkable properties, including these:

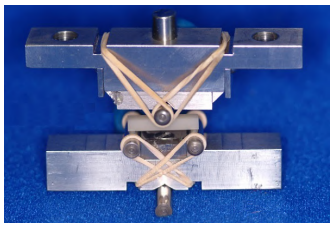
- It is the estimate of the measurand with smallest uncertainty;
- The probability distribution of $\widehat{\theta}$ (which is the value of a random variable because it is a function of the data) is approximately Gaussian, and the quality of the approximation improves as the number, n, of inputs increases;

MLE can be used whenever there is an explicit relationship between the true value of the quantity one wishes to estimate, and the parameters of the probability distribution of the data that is used for the purpose.

For example, when the replicated observations are from a Gaussian distribution, and the true value of the quantity of interest is the mean of this distribution. Likewise, in the example presented below, the quantity of interest (the mean tensile strength of alumina) is an explicit function of the two parameters of the Weibull distribution used to model replicated observations of the stress at which coupons of alumina break in a flexure test.

- The standard uncertainties and correlations of the components of $\widehat{\eta}$ can be computed approximately based on the matrix of second-order partial derivatives of $\ln L_x(\theta)$ with respect to the components of θ.

These properties, and the ease with which the MLE can be computed, make maximum likelihood estimation a very attractive, general purpose technique. We illustrate the calculation of the MLE, and the evaluation of the associated uncertainty, using as inputs 30 observations made under conditions of repeatability, of the rupture stress of alumina coupons in a 3-point flexure test.



Three-point flexural strength test of alumina coupon. Courtesy of George D. Quinn (Material Measurement Laboratory, NIST).

The model selected for the variability of the replicated determinations is the Weibull probability distribution, which has two parameters, generally called shape and scale, but that, in this context, are usually called the characteristic (or, nominal) strength σ_C, and the Weibull modulus m, respectively.[54] Note that, throughout this example, the Greek letter σ is used to denote stress (with the same units as pressure), not standard deviation.

[54] 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. doi:10.1016/j.dental.2009.09.006

Consistently with the notation used for the general description of the MLE above, we should then write $\theta = (m, \sigma_C)$. The measurand is the tensile strength $\eta = \sigma_C \Gamma(1 + 1/m)$, which is the mean of that Weibull distribution (and Γ is the gamma function).

The Weibull model may be characterized by saying that the rupture stress σ of an alumina coupon is such that it has the following probability density:

$$p(\sigma_i \,|\, m, \sigma_C) = \frac{m}{\sigma_C} \left(\frac{\sigma_i}{\sigma_C} \right)^{m-1} \mathrm{e}^{(-\sigma_i/\sigma_C)^m},$$

where the scale parameter σ_C and the shape parameter m are positive quantities.

σ/MPa				
307	407	435	455	486
371	409	437	462	499
380	411	441	465	499
393	428	445	466	500
393	430	445	480	543
402	434	449	485	562

Rupture stress for 30 alumina coupons in a 3-point flexure test. Courtesy of George D. Quinn (Material Measurement Laboratory, NIST).

Assuming that the $n = 30$ replicates are like outcomes of independent Weibull random variables, with $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ denoting the vector of observations, the likelihood function is L such that

$$L(m, \sigma_C \,|\, \sigma) = \prod_{i=1}^{n} p(\sigma_i \,|\, m, \sigma_C).$$

The maximum likelihood estimates of the Weibull modulus and of the characteristic strength are the values of m and σ_C that maximize $L(m, \sigma_C \,|\, \sigma)$ as a function of m and σ_C, with $\sigma_1, \ldots, \sigma_n$ kept fixed at the observed rupture stresses.

Since $L(m, \sigma_C \,|\, \sigma)$ is a product of terms involving m and σ_C, it is generally preferable to maximize $\ln L(m, \sigma_C \,|\, \sigma)$ instead. The reason is that the gradient of a sum is generally better behaved during numerical optimization than the gradient of a product, in the sense that the second derivatives either do not change too much or too rapidly. The

following R code minimizes $-\ln L(m, \sigma_C \mid \sigma)$, which is equivalent to maximizing the likelihood function.

```
sigma = c(307, 371, 380, 393, 393, 402, 407, 409, 411, 428,
          430, 434, 435, 437, 441, 445, 445, 449, 455, 462,
          465, 466, 480, 485, 486, 499, 499, 500, 543, 562)
negLogLik = function(par, s = sigma) {
  -sum(dweibull(s, shape=par[1], scale=par[2], log=TRUE)) }
# Find maximum likelihood estimates
opt = optim(par = c(m=10.6, sigmaC=465), fn = negLogLik,
             s = sigma, hessian = TRUE)
# Estimates of the shape and scale parameters
opt$par
# Approximate covariance matrix of the parameter estimates
V = solve(opt$hessian)
# Approximate standard uncertainties of the parameter estimates
sqrt(diag(V))
```

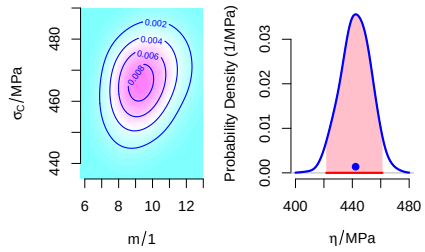
The R function `optim` minimizes the value of the function `negLogLik` with respect to its argument, the vector `par`, whose elements are the Weibull parameters, using the Nelder-Mead method [Nelder and Mead, 1965]. It requires that initial guesses be provided for the values of the parameters. The code requests that the matrix of second-order partial derivatives (Hessian matrix) be computed and returned because its inverse is an approximation to the covariance matrix of the parameter estimates. The larger the sample size, which is 30 in this case, the better the approximation.

The results are $\widehat{m} = 9.24$, $\widehat{\sigma}_C = 467\,\text{MPa}$, hence $\widehat{\eta} = 443\,\text{MPa}$. The last line of the previous R code will produce approximate evaluations of $u(\widehat{m}) = 1.23$ and $u(\widehat{\sigma}_C) = 9.8\,\text{MPa}$.

To compute $u(\widehat{\eta})$ one can use the fact that $\eta = \sigma_C \Gamma(1 + 1/m)$ is a measurement model of the kind the GUM contemplates, while recognizing that $\widehat{m}$ and $\widehat{\sigma}_C$ are correlated. The correlation between them is 0.33, which can be obtained in R using `cov2cor(V)`. The *NIST Uncertainty Machine* then yields $u(\widehat{\eta}) = 10.4\,\text{MPa}$.

These uncertainty evaluations all are made possible by the aforementioned MLE magic. However, this magic requires a large number of observations, while we have only 30. May this be enough? To answer this question, and to avoid this magic, we can redo the uncertainty analysis employing the parametric statistical bootstrap [Efron and Tibshirani, 1993], and compare the evaluations we will get this way with those we got above. The idea is to take the above MLEs of m and σ_C and use them to generate many samples of size 30 from the Weibull distribution with these values of the parameters. For each such sample, we find the best parameter values by minimizing $-\ln L(m, \sigma_C \mid \sigma)$.

```
m.HAT = opt$par['m']
sigmaC.HAT = opt$par['sigmaC']
boot = array(dim=c(1e5, 3))
colnames(boot) = c('m', 'sigmaC', 'eta')
for (j in 1:1e5) {
  sigmaB = rweibull(30, shape=m.HAT, scale=sigmaC.HAT)
  thetaB.MLE = optim( par = c(m=10, sigmaC=440),
                      fn = negLogLik, s = sigmaB)$par
  ## Calculate eta
  etaB = thetaB.MLE['sigmaC']*gamma(1 + 1/thetaB.MLE['m'])
  boot[j,] = c(thetaB.MLE, etaB)
}
apply(boot, 2, sd)
```



The contour lines in the left panel outline the shape of the joint probability density of $\widehat{m}$ and $\widehat{\sigma}_C$. The region shaded pink in the right panel amounts to 95 % of the area under the curve, hence its footprint on the horizontal axis is a 95 % coverage interval for the true value of η.

This R code produces $u(\widehat{m}) = 1.47$, $u(\widehat{\sigma_C}) = 9.8\,\text{MPa}$, and $u(\widehat{\eta}) = 10.5\,\text{MPa}$. Not only does this exercise validate the MLE magic in this case, it also gives us the ingredients fully to characterize the joint probability distribution of $\widehat{m}$ and $\widehat{\sigma}_C$, as well as the distribution of $\widehat{\eta}$.

Least Squares

Least squares is a criterion of estimation, often also described as a method for the adjustment of observations. Consider the simplest instance of such adjustment, where one has made m replicated determinations of the same quantity, $x_1, \ldots, x_m$, which one wishes to combine by choosing the value θ that minimizes the sum of squared deviations of the observations from it: $S(\theta) = (x_1 - \theta)^2 + \ldots + (x_m - \theta)^2$. Such θ is the solution of $S'(\theta) = 0$, where S' denotes the first derivative of S with respect to θ. That is, $(-2)(x_1 - \theta) + \ldots + (-2)(x_m - \theta) = 0$. Solving this equation for θ yields $\theta = (x_1 + \cdots + x_m)/m = \overline{x}$, the average of the observations. This is indeed the value where $S(\theta)$ achieves its minimum because $S''(\theta) = 2m > 0$.

If the measurement errors are Gaussian, then least squares is equivalent to maximum likelihood estimation. The method was developed by Adrien-Marie Legendre (1752–1833) and Carl Friedrich Gauss (1777–1855) at the beginning of the 19th century. In an early, and most remarkable application of the method, Gauss predicted where the asteroid Ceres should be found again after it had last been observed by its discoverer Giuseppe Piazzi.[55] And it was indeed at the location predicted by Gauss that Franz Xaver von Zach and Heinrich Olbers spotted Ceres in the skies on the last day of 1801.

[55] C. F. Gauss. Summarische Uberficht der zur bestimmung der bahnen der beyden neuen hauptplaneten augewanden methoden. *Monatliche Correspondenz zur Beförderung der Erd- und Himmels-Kunde*, XX(Part B, July-December, Section XVII):197–224, September 1809

The method of least squares can be illustrated with an example we encountered earlier — determining the mass of three objects whose mass differences were recorded using a mass comparator. This example involves three observations ($D_{AB} = -0.38\,\text{mg}$, $D_{AC} = -1.59\,\text{mg}$, and $D_{BC} = -1.22\,\text{mg}$), three parameters whose values we are seeking (δ_A, δ_B, and δ_C), and a constraint $K = \delta_A + \delta_B = 0.83\,\text{mg}$ that must be satisfied while also taking into account its associated uncertainty, $u(K) = (0.07\,\text{mg}) \cdot \sqrt{2}$.

If measurement errors are best modeled using a probability distribution other than Gaussian, then an adjustment of observations based on a different criterion may be preferable. For example, minimizing the sum of the absolute values of the errors will lead to the *median*, which is the maximum likelihood solution when the errors follow a Laplace (double -exponential) distribution.

The three observations are mutually inconsistent because, for example, $D_{AB} - D_{AC} = -1.21\,\text{mg}$ while $D_{BC} = -1.22\,\text{mg}$. To make them consistent we introduce non-observable "errors" ε_1, ε_2, and ε_3, such that the following three equations hold true simultaneously

$$\begin{aligned} D_{AB} &= \delta_A - \delta_B + \varepsilon_1, \\ D_{AC} &= \delta_A - \delta_C + \varepsilon_2, \\ D_{BC} &= \delta_B - \delta_C + \varepsilon_3. \end{aligned}$$

Applying the method of least squares in this case amounts to choos-

ing values for δ_A, δ_B, and δ_C that minimize the sum of the squared errors, $\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2$, while also satisfying the constraint

$$K = \delta_A + \delta_B = 0.83\,\text{mg}.$$

This constraint is "soft" because it is surrounded by uncertainty, $u(K) = (0.07\,\text{mg}) \cdot \sqrt{2}$. However, let us begin by pretending that it is "hard" so that we can replace δ_B with $K - \delta_A$ and write the optimization criterion as follows:

$$\begin{aligned} S(\delta_A, \delta_C) &= \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^3 \\ &= (D_{AB} - \delta_A + (K - \delta_A))^2 + (D_{AC} - \delta_A + \delta_C)^2 \\ &+ (D_{BC} - (K - \delta_A) + \delta_C)^2. \end{aligned}$$

The values of δ_A and δ_C that minimize $S(\delta_A, \delta_C)$ correspond to a situation when both partial derivatives equal zero, $\partial S(\delta_A, \delta_C)/\partial\delta_A = 0$ and $\partial S(\delta_A, \delta_C)/\partial\delta_C = 0$, that is

$$\begin{aligned} \widehat{\delta_A} &= D_{AB}/3 + D_{AC}/6 - D_{BC}/6 + K/2 = 0.227\,\text{mg} \\ \widehat{\delta_C} &= -D_{AC}/2 - D_{BC}/2 + K/2 = 1.82\,\text{mg}. \end{aligned}$$

These indeed correspond to a minimum of the criterion because the matrix of second order partial derivatives of $S(\delta_A, \delta_C)$ is diagonal and both elements in its main diagonal are positive. Applying the constraint yields the estimate of the remaining parameter, $\widehat{\delta_B} = K - \widehat{\delta_A} = 0.603\,\text{mg}$.

Now we need to bring into play the "softness" of the constraint, which is its uncertainty, $u(K)$. This can be accomplished in any one of several different ways. The most intuitive one may be a Monte Carlo procedure.

The idea is to solve the same optimization problem we just solved, when we pretended that the constraint was "hard", but to do it many times over, each time using a value for the constraint drawn from a probability distribution with mean K and standard deviation $u(K)$. We will use a Gaussian distribution for this purpose, in keeping with the spirit of least squares.

```
D.AB = -0.38; D.AC = -1.59; D.BC = -1.22
abc = array(dim=c(1e6, 3))
for (i in 1:1e6) {
  k = rnorm(1, mean=0.83, sd=0.07*sqrt(2))
  A = D.AB/3 + D.AC/6 - D.BC/6 + k/2
  B = k - a
  C = -D.AC/2 - D.BC/2 + k/2
  abc[i,] = c(A, B, C)
}
apply(abc, 2, mean)
apply(abc, 2, sd)
```

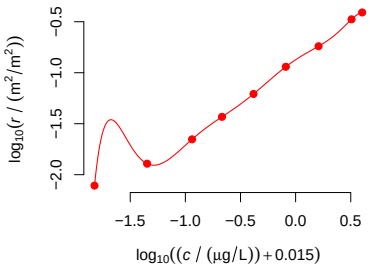
The final, constrained least squares estimates are $\widehat{\delta}_A = 0.227\,\text{mg}$, $\widehat{\delta}_B = 0.603\,\text{mg}$, and $\widehat{\delta}_C = 1.82\,\text{mg}$, with associated uncertainties $u(\widehat{\delta}_A) = 0.049\,\text{mg}$, $u(\widehat{\delta}_B) = 0.049\,\text{mg}$, and $u(\widehat{\delta}_C) = 0.049\,\text{mg}$.

More general constrained least squares problems can be solved using the method of Lagrange multipliers, as described by Zelen [1962] and Seber [2008, §24.3]. R function `solnp`, in package `Rsolnp` implements a versatile algorithm for constrained, nonlinear optimization using an augmented Lagrange method.[56]

[56] Y. Ye. *Interior Point Algorithms: Theory and Analysis*. John Wiley & Sons, New York, NY, 1997. ISBN 978-0471174202; and A. Ghalanos and S. Theussl. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, 2015. R package version 1.16

The method of least squares is very often used to fit models to data, and it is also very often misused because users fail to realize how attentive this method is toward every little detail in the data, while such solicitude may, in many cases, prove excessive. For example, a single data point that markedly deviates from the pattern defined by the others can lead the least squares fit astray, and a least squares fit may reproduce the data exactly yet be ridiculous.

A figure presented earlier and reproduced here illustrates this point in spades. The fit, which may be computed using the R code below, goes through each data point exactly, but at the price of an odd, obviously unrealistic contortion of the curve. The residuals, which are the differences between observed and fitted values of $\log_{10}(r/(\text{m}^2/\text{m}^2))$, are all zero because the method of least squares forces a polynomial (regardless of degree), with as many coefficients as there are data points, to pass through all of them, at any cost.



Even though a polynomial of the 8th degree fits the median values of r at each value of c (red dots) exactly, it would be an unrealistic calibration function.

```
x = c(-1.824, -1.347, -0.939, -0.668, -0.382, -0.089, 0.208, 0.507, 0.604)
y = c(-2.107, -1.892, -1.653, -1.432, -1.208, -0.942, -0.74, -0.476, -0.409)
summary(lm(y~poly(x, degree=8, raw=TRUE)))
```

When the method of least squares is used either to adjust observations, or to fit a function to empirical data, often it is applied subject to constraints. For example, when the purpose is to adjust mass fractions of a compound whose constituents are determined separately from one another, one will wish to constrain the adjusted mass fractions to be non-negative, or to be less than 1 g/g, or to sum to 1 g/g, or possibly to satisfy more than one such constraint simultaneously. Similarly, when fitting a piecewise polynomial function to data, one may wish to constrain the result to be continuous and smooth, that is, to be a *spline* [Ferguson, 1986].

Between approximately −260 °C and 960 °C, the International Temperature Scale (ITS-90) is defined by means of platinum resistance thermometers calibrated at specified fixed points (such as melting points of various metals). In this temperature interval, the ITS-90 reference function is given by two high-order polynomials constrained to join at the triple point of water without discontinuity of either the polynomials or of their first derivatives.

Model Selection

When we built a model for the calibration function used to measure the mass concentration of chloromethane we employed the Bayesian Information Criterion (BIC) as a guide to select one among several alternative models, and pointed out that the smaller the BIC, the more adequate the model. Here we describe how BIC is computed,

and explain why the best model (among several under consideration) has the smallest value of BIC.

Consider fitting a straight line (first degree polynomial) to the same data we considered above, using the following R code:

```
x = c(-1.824, -1.347, -0.939, -0.668, -0.382, -0.089, 0.208, 0.507, 0.604)
y = c(-2.107, -1.892, -1.653, -1.432, -1.208, -0.942, -0.74, -0.476, -0.409)
summary(lm(y~poly(x, degree=1, raw=TRUE)))
```

The model treats the $m = 9$ values of x as known without uncertainty, and regards the values of y as outcomes of m independent Gaussian random variables whose means depend on the values of x. More precisely, y_i is an outcome of a Gaussian random variable with mean $\beta_0 + \beta_1 x_i$ and standard deviation σ, for $i = 1, \ldots, m$.

The *likelihood function* corresponding to these data is a function L of the three parameters β_0, β_1, and σ, where the data $\{x_i, y_i\}$ are kept fixed, such that

$$L(\beta_0, \beta_1, \sigma \,|\, x, y) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^m \exp\left\{-\sum_{i=1}^{m} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\}.$$

In these circumstances, the maximum likelihood estimates of β_0, and β_1 are the least squares estimates, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, and the maximum likelihood estimate of σ^2 is the average of the squared residuals $\{y_i - \widehat{y}_i\}$, where $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, for $i = 1, \ldots, m$: that is,

$$\widehat{\sigma}^2 = \sum_{i=1}^{m} (y_i - \widehat{y}_i)^2 / m.$$

The BIC for this model and data is

$$\text{BIC} = -2 \ln L(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma} \,|\, x, y) + k \ln m,$$

where $k = 3$ denotes the number of model parameters. The closer the model fits the data, the larger the value $L(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma} \,|\, x, y)$ that the likelihood function takes at the maximum likelihood estimates. Or, equivalently, the more accurate the model, the smaller (the more negative) the first term on the right-hand side of the foregoing definition of the BIC will be (because it has a minus sign in front of it and the logarithm is an increasing function).

In general, the larger the number of parameters in a model, the closer it will fit the data. Therefore, the greater the degree of the polynomial, the closer the fit (above we saw that a polynomial of the 8th degree will fit these data exactly), and the smaller (the more negative) $-\ln L$ will be. On the other hand, since k denotes the number of parameters in the model, the larger this number the larger the second term in the definition of the BIC, which is added to the first.

That is, the two terms in the BIC move in opposite directions as the number of parameters in the model increases: the first term becomes

Note that k is not the degree of the polynomial; it is the number of adjustable parameters. For polynomial regression models, like the ones we are comparing here, k is the number of coefficients of the polynomial plus the additional parameter, σ.

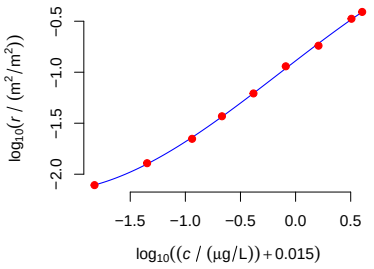
DEGREE	k	BIC
1	3	−19.9
2	4	−32.3
3	5	−43.4
4	6	−41.3
5	7	−43.1
6	8	−41.7
7	9	−41.3

The smaller the value of the Bayesian information criterion, BIC, the more adequate the model for the data. In this case BIC decreases appreciably as the degree of the polynomial increases from 1 to 3, but then stabilizes, fluctuating around the same value. This suggests that a polynomial of the third degree may be the best choice for these data.

smaller, while the second increases. The first term rewards goodness of fit (the smaller the better), while the second term, $k \ln m$, penalizes model complexity (the larger the worse), where "complexity" here means number of adjustable parameters. In summary, when we select the model that minimizes BIC we are striking a compromise between goodness-of-fit and model complexity.

The following R code computes BIC for the data and first degree polynomial model described above. It does it both from scratch and also using the built-in function `BIC`.

```
x = c(-1.824, -1.347, -0.939, -0.668, -0.382, -0.089, 0.208, 0.507, 0.604)
y = c(-2.107, -1.892, -1.653, -1.432, -1.208, -0.942, -0.74, -0.476, -0.409)
y1.lm = lm(y~poly(x, degree=1, raw=TRUE))
n = nrow(y1.lm$model) ## Size of the sample the model was fitted to
k = length(y1.lm$coefficients) + 1 ## sigma is the extra parameter
sigmaHAT = sqrt(mean(residuals(y1.lm)^2)) ## MLE of sigma
yHAT = fitted.values(y1.lm)
loglik = sum(dnorm(y, mean=yHAT, sd=sigmaHAT, log=TRUE))
c(BIC=-2*loglik + k*log(n), BIC=BIC(y1.lm))
```



The best model according to BIC, among those under consideration, is a polynomial of the 3rd degree.

Bayesian Estimation

Bayesian estimation provides the means to blend prior information about the measurand with the fresh information in newly acquired measurement results.[57]

The prior information may originate in similar studies carried out in the past, or it may reflect expert knowledge: in either case, it must be cast in the form of a probability distribution on the set of possible values of the measurand. When an expert is the source of prior information, one should employ a disciplined approach to elicit the relevant information and to encapsulate it in a probability distribution.[58]

Besides this practical value, the Bayesian approach to drawing inferences from data also aligns the interpretation of such inferences with how most people are naturally inclined to interpret them. This advantage is clearest in relation with the interpretation of coverage intervals.

The conventional interpretation, which has pervaded the teaching of statistics for at least 70 years now, goes like this: a 95 % interval for the true value of a quantity is a realization of a random interval, and the 95 % probability does not apply specifically to the interval one actually gets, but is a property of the procedure that generates such interval.

This interpretation typically goes hand in hand with an interpretation of probability that equates it with frequency in the long run. In this vein, one finds statements like this: the 95 % means that, of all such intervals that a statistician produces in her lifetime, 95 % cover

[57] A. O'Hagan. The Bayesian Approach to Statistics. In T. Rudas, editor, *Handbook of Probability: Theory and Applications*, chapter 6. Sage Publications, Thousand Oaks, CA, 2008. ISBN 978-1-4129-2714-7. doi:10.4135/9781452226620.n6

The Bayesian approach is also very useful in situations where there are about as many parameters as there are observations, for example in image reconstruction [Geman and Geman, 1984].

[58] 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; and 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, February 2014. doi:10.1016/j.envsoft.2013.10.010

their intended targets, and 5 % miss them.

The Bayesian interpretation of coverage intervals is more intuitive, and certainly applies to the specific interval that one actually gets: the 95 % is the probability that the value of interest is in that particular interval that one has computed.

Bayesian statistics gets its name from the 18th century English statistician, philosopher, and minister, Thomas Bayes, whose most famous accomplishment was published only posthumously [Bayes and Price, 1763].

This interpretation is enabled by a change in viewpoint: the interval one gets is as concrete and definite as can be — there being nothing random about it. The "randomness" is transferred to the quantity whose true value is unknown, while the very meaning of "random" is refreshed. From a Bayesian viewpoint, a random quantity does not have a value that fluctuates unpredictably like a leaf fluttering in the wind — its value is what it is, and either we just do not know it at all, or our knowledge of it is incomplete.

Bayesians use probability distributions to quantify degrees of belief (in the truth of propositions about the true values of properties under study), or to describe states of partial or incomplete knowledge about these properties. A random variable is simply a property (quantitative or qualitative) that has a probability distribution as an attribute. This attribute is not an intrinsic attribute of the property. Instead, it describes an epistemic relation between the person aiming to learn the true value of the property, and this true value.

The Bayesian approach is eminently practical because its specific results have the meaning common sense expects them to have, and they are immediately relevant because they are not contingent on what may happen in the rest of anyone's lifetime (refer to the discussion above of the meaning of confidence intervals).

In a nutshell, the Bayesian approach to estimation and uncertainty evaluation for statistical measurement models involves modeling all parameters whose true values are unknown as (non-observable) values of random variables, and the measurement data as observed outcomes of random variables whose distributions depend on the unknown parameters values. The estimate of the measurand, and an evaluation of the associated uncertainty, are derived from the conditional distribution of the unknowns given the data.

We demonstrate the approach in the context of the estimation of the tensile strength η of alumina coupons, which we addressed above using the method of maximum likelihood estimation.

The *prior distribution* tells us the likely whereabouts of the parameters before we gather any new data. The *likelihood* tells us how likely the data are given any particular values of the parameters. Bayes's rule [DeGroot and Schervish, 2012, 2.3,7] puts these two pieces together to tell us how likely it is that the true values of the parameters will be in any specified subsets of their ranges, in light of the fresh data, and with due allowance for the prior information.

(a) The prior knowledge in hand consists of facts about the Weibull modulus m and the characteristic strength σ_C that have been established in previous studies of rupture of the same material, also in 3-point flexure testing: that m is around 8.8, give or take 1.25, and that σ_C is around 467 MPa give or take 11 MPa. We capture these facts by modeling m and σ_C *a priori* as independent

random variables with Gaussian distributions, m with mean 8.8 and standard deviation 1.25, σ_C with mean 467 MPa and standard deviation 11 MPa. This defines the *prior distribution*, whose probability density, π, is the product of two Gaussian probability densities, one for m, the other for σ_C.

(b) Given any hypothetical values of m and σ_C, the observed values of rupture stress, for the 30 coupons that were tested, are modeled as outcomes of 30 independent random variables, all with the same Weibull distribution with shape m and scale σ_C. The product of the corresponding 30 Weibull densities, each evaluated at an observed value of rupture stress, then becomes a function of m and σ_C alone (the observations of rupture stress, $\{\sigma_i\}$, are all frozen at their observed values). This is the same likelihood function, $L_{\boldsymbol{\sigma}}(m, \sigma_C)$, where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{30})$, that we encountered while discussing maximum likelihood estimation.

(c) The conditional distribution of the parameters given the data (which actually is the version of the prior distribution suitably updated by incorporation of the fresh data), the so-called *posterior distribution*, has probability density given by Bayes's Rule:

$$q_{\boldsymbol{\sigma}}(m, \sigma_C) = \frac{L_{\boldsymbol{\sigma}}(m, \sigma_C) \cdot \pi(m, \sigma_C)}{\int_0^{+\infty} \int_0^{+\infty} p_{s,t}(\boldsymbol{\sigma}) \pi(s, t) \mathrm{d}s\mathrm{d}t}.$$

Referring to Bayes's Rule, Jeffreys [1973, 2.3] pointed out that "This theorem (due to Bayes) is to the theory of probability what Pythagoras's theorem is to geometry."

Typically, Bayes's Rule is not used directly in practice because the formula that it produces for the probability density of m and σ_C given the data and the prior information involves integrals that cannot be evaluated analytically, and that may be impracticable to compute numerically. Other tools have to be employed to coax the wheels of the Bayesian machinery to turn.

An invention dating back to the 1950's, Markov Chain Monte Carlo (MCMC) sampling[59], coupled with the contemporary prevalence of fast personal computers, has revolutionized the practice of Bayesian statistics.

[59] C. Robert and G. Casella. A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26(1):102–115, 2011. doi:10.1214/10-STS3510

MCMC frees users from constraints of mathematical tractability, and allows them to employ realistically appropriate Bayesian models and still be able to draw samples from the posterior distribution without computing its density explicitly (for example, $q_{\boldsymbol{\sigma}}$ above).

The Russian mathematician Andrey Markov (1856-1922) found that the sequence of consonants and vowels in Alexander Pushkin's *Eugene Onegin* could be described as a random sequence with a particular structure: the probability of the appearance of a vowel or consonant largely depends only on the type of letter immediately preceding it. This model is still in use today to help identify the authors of texts of unknown authorship [Khmelev and Tweedie, 2001].

Markov Chain Monte Carlo is an iterative procedure. At each step, first it generates proposed values for the parameters by making random drawings from a suitable (generally multivariate) distribution (fittingly called the *proposal distribution*).

Then it compares the proposed values with the values that had been accepted in the previous step by computing the ratio r of the value that the posterior probability density takes at the proposed parameter values, to the value it takes at the previously accepted parameter values. (Note that, to compute this ratio only the numerator of Bayes's formula needs to be evaluated, not the denominator, which usually is the challenging or impracticable piece to compute.)

When $r > 1$, the proposed values of the parameters are accepted for the current step without further ado. When $r \leqslant 1$, a number z is drawn that is distributed uniformly between 0 and 1: if $z < r$, then the proposed values are still accepted; otherwise, the proposal is rejected and the values of the parameters in the previous step are taken also for the current step.

Since the result of each step depends only on the result of the previous step, the resulting sequence of parameter values is a Markov chain on the space of parameter values. The manner, specified above, of transitioning from one step to the next, ensures that the stationary (or, equilibrium) distribution of this Markov chain is the posterior probability distribution sought.

The chain eventually "forgets" its initial state — which is an arbitrary assignment of values to the parameters —, and the sequence of accepted values of the parameters is like a sample from the posterior distribution, albeit with some dependence.

Nowadays there are many different ways of implementing MCMC. The procedure sketched above is one of the oldest, called the *Metropolis-Hastings algorithm* [Metropolis et al., 1953; Hastings, 1970].

The following R code shows an example of how the Markov Chain Monte Carlo can be used to sample the joint posterior distribution of m and σ_C. We begin by defining an R function that computes the logarithm of the numerator of Bayes's Rule.

```
lup = function (theta, x) {
    m = theta[1]; sigmaC = theta[2]
    ## Prior distribution for m
    prior.m = dnorm(m, mean=8.8, sd=1.25, log=TRUE)
    ## Prior distribution for sigmaC
    prior.s = dnorm(sigmaC, mean=467, sd=11, log=TRUE)
    ## Log-likelihood function
    loglik = sum(dweibull(x, shape=m, scale=sigmaC, log=TRUE))
    ## Compute value of numerator of Bayes rule by summing
    ## the logarithms of the prior densities and of the likelihood
    return(prior.m + prior.s + loglik) }
```

The R function `lup` evaluates the logarithm of the numerator of Bayes's rule, $\ln(p_{m,\sigma_C}) + \ln(\pi)$, which is all that is needed to be able to do MCMC. (The name "`lup`" refers to the logarithm of the unnormalized posterior density.)

Next we place the determinations of rupture stress that we used above, when discussing maximum likelihood estimation, into the vector `sigma`, and then take K steps of the Markov chain defined above, drawing candidate values for the parameters from Gaussian distributions. Since some of these values could conceivably be negative, we effectively truncate the proposal distributions at zero, thus ensuring that the candidate values for the parameters are always positive, which they must be because they are the shape and scale of a Weibull distribution.

Once these K steps are completed, we discard the initial 25 % of the chain to remove the effects of the starting values, 9 and 470 MPa. And we keep only every 20th pair of parameter values from the remaining steps to reduce the impact that correlations between accepted values may have upon the estimates of standard uncertainty for the Bayes estimates that we will derive from the MCMC sample.

```
## Determinations of rupture stress of alumina coupons (MPa)
sigma = c(307, 371, 380, 393, 393, 402, 407, 409, 411, 428,
          430, 434, 435, 437, 441, 445, 445, 449, 455, 462,
          465, 466, 480, 485, 486, 499, 499, 500, 543, 562)
K = 1e6
mcmc = array(dim=c(K,2))
## Assign initial values to Weibull parameters
mcmc[1,] = pars = c(m=9, sigmaC=470)
for (k in 2:K) {
    ## Generate new candidate values for the parameters
    ## in the vicinity of the previous values,
    ## while ensuring that both are positive because they are
    ## supposed to be Weibull shape and scale parameters
    parsCandidate = abs(rnorm(2, mean=pars, sd=0.05*pars))
    ## Calculate the acceptance ratio
    r = exp(lup(parsCandidate, x=sigma) - lup(pars, x=sigma))
    ## Accept candidate values if r is greater than
    ## a number drawn uniformly at random from [0,1]
    if (r > runif(1)) { mcmc[k,] = parsCandidate
        pars = parsCandidate } else { mcmc[k,] = pars }
}
## Discard the initial 25 percent of the chain,
## and keep only every 20th of the accepted parameters
mcmc = mcmc[seq(0.75*nrow(mcmc), nrow(mcmc), by=20),]
m.TILDE = mcmc[,1]
sigmaC.TILDE = mcmc[,2]
```

(d) What do we do with such sample? The sky is the limit, really, because by making this sample very large (which can be done at the expense of very quick computation), we characterize it sufficiently well to be able to compute any function of it that will be required, and to do so with high accuracy.

In the case we are considering, this sample comprises pairs of values of m and σ_C (which, *a posteriori*, are no longer indepen-

dent, because they draw information from the same data). The first thing we do with this sample of pairs of values of the parameters is to compute a value of η from each of these pairs, thus producing a sample from the distribution of the measurand.

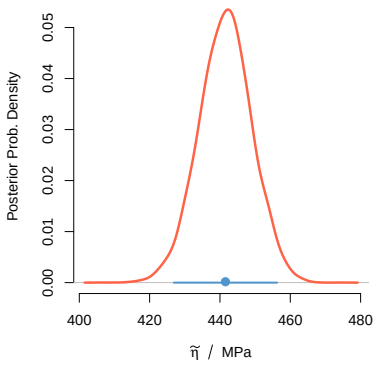
```
eta.TILDE = sigmaC.TILDE*gamma(1 + 1/m.TILDE)
```

Then we can reduce this sample in any way that is fit for purpose: by computing its mean or its median, its standard deviation, coverage intervals of any probability, etc.

The MLE and Bayes estimates of η, 443 MPa and 442 MPa, are almost identical, but the associated uncertainties are markedly different: MLE's is 10.4 MPa, while its Bayesian counterpart is 7.8 MPa.

The estimates are almost identical because the information in the data is in very close agreement with the prior information, and because there is enough data to weigh fairly heavily upon any possibly unfortunate specification of prior information. The uncertainty for the Bayes estimate is appreciably smaller than for the MLE because the prior information is very specific, which the MLE is not privy to. In fact, the MLE may be interpreted as a particular Bayesian estimate (the so-called maximum *a posteriori* estimate) when the parameters are uniformly distributed *a priori* over their ranges (such uniform distribution is a concept of questionable validity, considering that their ranges are infinitely long).

The power of Bayesian methods lies in the fact that they allow us to incorporate relevant information that the likelihood function may be unable to accommodate. For example, natural constraints that the parameter values must satisfy, or information about the precision of some parameters.



Posterior probability density of η obtained using the simple implementation of the MCMC sampler described above, posterior mean (blue dot), and 95 % credible interval centered at the posterior mean (blue line segment).

THE MASS FRACTION OF NITRITE IONS in a sample of seawater was measured using Griess's method,[60] based on four determinations obtained under conditions of repeatability:

$$w(\mathrm{NO_2^-}) = 0.1514, 0.1523, 0.1545, 0.1531 \text{ mg/kg}$$

While we might not have any strong prior information about the nitrite levels in this seawater sample, based on the performance of the measurement method we do expect that the relative measurement uncertainty is 1 % to within a factor of 3. We can model this prior knowledge about the standard deviation, σ, of the measurement errors affecting the individual determinations, using a gamma distribution whose 10th and 90th percentiles are 0.33 % and 3 % of 0.150 mg/kg, respectively. Using R we can obtain the parameters of the gamma distribution that has these percentiles as follows:

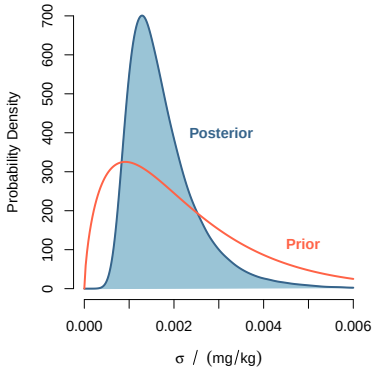
[60] P. Griess. Bemerkungen zu der abhandlung der hh. weselsky und benedikt "ueber einige azoverbindungen". *Berichte der Deutschen Chemischen Gesellschaft*, 12(1):426–428, 1879. doi:10.1002/cber.187901201117

```
require(rriskDistributions)
get.gamma.par(p = c(0.10, 0.90), q = 0.150*c(1/3, 3)/100)
```

This yields shape $\alpha = 1.696$ and rate $\beta = 762.3\,\mathrm{kg/mg}$.

The following Stan and R codes fit the model $w_i(\mathrm{NO_2^-}) = \omega + \varepsilon_i$ to the replicate determinations $i = 1, 2, 3, 4$, where ω denotes the true value of the mass fraction of nitrite in the sample of seawater, and the measurement errors $\{\varepsilon_i\}$ are assumed to be a sample from a Gaussian distribution with mean 0 and standard deviation σ. The prior information about σ is encapsulated in the gamma distribution specified above. For ω we adopt a weakly informative Gaussian prior distribution.

```
require(rstan)
w = c(0.1514, 0.1523, 0.1545, 0.1531)
m = "data { real w[4]; }
     parameters {
       real<lower=0> omega;
       real<lower=0> sigma;
     }
     model {
       // Priors on parameters
       // True mean mass fraction of nitrite
       omega ~ normal(0, 1);
       // Std. Dev. of measurement errors
       sigma ~ gamma(1.696, 762.3);
       // Likelihood
       w ~ normal(omega, sigma);
     }"
fit = stan(model_code = m, data = list(w=w),
           warmup=75000, iter=750000,
           chains=4, cores=4, thin=25)
print(fit, digits=5)
```



Prior and posterior probability densities for σ. The relative prior uncertainty about σ, which is 77 %, is reduced to 47 % after incorporation of the observations.

The posterior mean of ω is 0.1528 mg/kg, with standard uncertainty 0.0010 mg/kg, which is 50 % larger than the conventional Type A evaluation of the standard uncertainty for the average of the replicates.

Bibliography

N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters. arXiv:1807.06209, 2018. URL https://arxiv.org/abs/1807.06209.

A. Agresti. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, 3rd edition, 2019.

C. Ainsworth. Sex redefined. *Nature*, 518:288–291, February 2015. doi:10.1038/518288a. News Feature.

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. doi:10.2307/2987937.

D. G. Altman and J. M. Bland. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *British Medical Journal*, 308(6943):1552, 1994. doi:10.1136/bmj.308.6943.1552.

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. doi:10.1214/aoms/1177729437.

G. Audi, F.G. Kondev, M. Wang, W. J. Huang, and S. Naimi. The NUBASE2016 evaluation of nuclear properties. *Chinese Physics C*, 41 (3):030001–1–138, March 2017. doi:10.1088/1674-1137/41/3/030001.

W. Bablok, H. Passing, R. Bender, and B. Schneider. A general regression procedure for method transformation. application of linear regression procedures for method comparison studies in clinical chemistry, part iii. *Clinical Chemistry and Laboratory Medicine*, 26: 783–790, 1988. doi:10.1515/cclm.1988.26.11.783.

A. Baddeley and R. Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42, 2005. URL www.jstatsoft.org/v12/i06/.

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.

Mr. Bayes and Mr. Price. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, January 1763. doi:10.1098/rstl.1763.0053. Communicated by Mr. Price, in a letter to John Canton.

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

R. P. Binzel. Pluto-Charon mutual events. *Geophysical Research Letters*, 16(11):1205–1208, November 1989. doi:10.1029/gl016i011p01205.

S. Birrer et al. H0LiCOW – IX. Cosmographic analysis of the doubly imaged quasar SDSS 1206+4332 and a new measurement of the Hubble constant. *Monthly Notices of the Royal Astronomical Society*, 484(4):4726–4753, January 2019. doi:10.1093/mnras/stz200.

O. N. Bjørnstad. *Epidemics — Models and Data using R*. Springer, Cham, Switzerland, 2018. ISBN 978-3-319-97486-6. doi:10.1007/978-3-319-97487-3.

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. doi:10.1177/096228029900800204.

BMJ News and Notes. Epidemiology — influenza in a boarding school. *British Medical Journal*, 1:586–590, March 1978. doi:10.1136/bmj.1.6112.586.

B. Braden. The surveyor's area formula. *The College Mathematics Journal*, 17(4):326–337, 1986. doi:10.2307/2686282.

H. Burgess and B. Spangler. Consensus building. In G. Burgess and H. Burgess, editors, *Beyond Intractability*. Conflict Research Consortium, University of Colorado, Boulder, Colorado, USA, September 2003. URL www.beyondintractability.org/essay/consensus-building.

P.-C. Bürkner. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411, 2018. doi:10.32614/RJ-2018-017.

K. P. Burnham and D. R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304, November 2004. doi:10.1177/0049124104268644.

S. G. Burrard. Mount Everest: The story of a long controversy. *Nature*, 71:42–46, November 1904. doi:10.1038/071042a0.

A. Canty and B. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2020. URL `cran.r-project.org/web/packages/boot/`. R package version 1.3-25.

B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1): 1–32, 2017. doi:10.18637/jss.v076.i01.

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

B. Carstensen, L. Gurrin, C. T. Ekstrøm, and M. Figurski. *MethComp: Analysis of Agreement in Method Comparison Studies*, 2020. URL `https://CRAN.R-project.org/package=MethComp`. R package version 1.30.0.

J. Chambers, W. Cleveland, B. Kleiner, and P. Tukey. *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA, 1983.

F. Chayes. On correlation between variables of constant sum. *Journal of Geophysical Research*, 65(12):4185–4193, 1960. doi:10.1029/JZ065i012p04185.

G. C.-F. Chen et al. A SHARP view of H0LiCOW: H_0 from three time-delay gravitational lens systems with adaptive optics imaging. *Monthly Notices of the Royal Astronomical Society*, 490(2):1743–1773, September 2019. doi:10.1093/mnras/stz2547.

M. R. Chernick. *Bootstrap Methods: A Guide for Practitioners and Researchers*. John Wiley & Sons, Hoboken, NJ, second edition, 2008. ISBN 978-0-471-75621-7.

Y. Cho, L. Hu, and H. *et al.* Hou. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nature Communications*, 4:2433, September 2013. doi:10.1038/ncomms3433.

W. S. Cleveland, E. Grosse, and W. M. Shyu. Local regression models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, chapter 8. Wadsworth & Brooks/Cole, Pacific Grove, California, 1992.

W. G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10(1):101–129, March 1954. doi:10.2307/3001666.

S. Cowley and J. Silver-Greenberg. These Machines Can Put You in Jail. Don't Trust Them. *The New York Times*, November 3, 2019. Business Section.

P. E. Damon, D. J. Donahue, B. H. Gore, A. L. Hatheway, A. J. T. Jull, T. W. Linick, P. J. Sercel, L. J. Toolin, C. R. Bronk, E. T. Hall, R. E. M. Hedges, R. Housley, I. A. Law, C. Perry, G. Bonani, S. Trumbore, W. Woelfli, J. C. Ambers, S. G. E. Bowman, M. N. Leese, and M. S. Tite. Radiocarbon dating of the Shroud of Turin. *Nature*, 337: 611–615, February 1989. doi:10.1038/337611a0.

A. C. Davison and D. Hinkley. *Bootstrap Methods and their Applications*. Cambridge University Press, Cambridge, UK, 1997. ISBN 0-521-57471-4. URL `statwww.epfl.ch/davison/BMA/`.

M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison-Wesley, Boston, MA, 4th edition, 2012.

P. Diaconis and B. Efron. Computer-intensive methods in statistics. *Scientific American*, 248:116–130, 1983.

A. Domínguez, R. Wojtak, J. Finke, M. Ajello, K. Helgason, F. Prada, A. Desai, V. Paliya, L. Marcotulli, and D. H. Hartmann. A new measurement of the Hubble Constant and matter content of the universe using extragalactic background light γ-ray attenuation. *The Astrophysical Journal*, 885(2):137, November 2019. doi:10.3847/1538-4357/ab4a0e.

R. L. Duncombe and P. K. Seidelmann. A history of the determination of pluto's mass. *Icarus*, 44:12–18, 1980. doi:10.1016/0019-1035(80)90048-2.

K. Dutta, A. Roy, Ruchika, A. A. Sen, and M. M. Sheikh-Jabbari. Cosmology with low-redshift observations: No signal for new physics. *Physical Review D*, 100:103501, November 2019. doi:10.1103/PhysRevD.100.103501.

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

D. R. Ferguson. Construction of curves and surfaces using numerical optimization techniques. *Computer-Aided Design*, 18(1):15–21, 1986. doi:10.1016/S0010-4485(86)80004-5.

R. A. Fisher. *Statistical Methods for Research Workers*. Hafner Publishing Company, New York, NY, 14th edition, 1973.

M. A. Fligner and T. J. Killeen. Distribution-free two-sample tests for scale. *Journal of the American Statistical Association*, 71(353):210–213, March 1976. doi:10.2307/2285771.

D. Freedman, R. Pisani, and R. Purves. *Statistics*. W. W. Norton & Company, New York, NY, 4th edition, 2007. ISBN 978-0-393-92972-0.

W. L. Freedman et al. The Carnegie-Chicago Hubble Program. VIII. an independent determination of the Hubble Constant based on the Tip of the Red Giant Branch. *The Astrophysical Journal*, 882(1):34, August 2019. doi:10.3847/1538-4357/ab2f73.

L. M. Friedman, C. D. Furberg, D. DeMets, D. M. Reboussin, and C. B. Granger. *Fundamentals of Clinical Trials*. Springer, Switzerland, 5th edition, 2015.

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/LM.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. URL `http://gdz.sub.uni-goettingen.de`.

C. F. Gauss. Summarische Uberficht der zur bestimmung der bahnen der beyden neuen hauptplaneten augewanden methoden. *Monatliche Correspondenz zur Beförderung der Erd- und Himmels-Kunde*, XX(Part B, July-December, Section XVII):197–224, September 1809.

A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–533, 2006. doi:10.1214/06-BA117A.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

A. Ghalanos and S. Theussl. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, 2015. R package version 1.16.

G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, and M. Colaneri. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, 26:855–860, June 2020. doi:10.1038/s41591-020-0883-7.

J. D. Giorgini. Status of the JPL Horizons Ephemeris System. In *IAU General Assembly*, volume 29, page 2256293, August 2015. URL `https://ssd.jpl.nasa.gov/`.

P. G. Gottschalk and J. R. Dunn. The five-parameter logistic: A characterization and comparison with the four-parameter logistic. *Analytical Biochemistry*, pages 54–65, 2005. doi:10.1016/j.ab.2005.04.035.

P. Griess. Bemerkungen zu der abhandlung der hh. wesel-sky und benedikt "ueber einige azoverbindungen". *Berichte der Deutschen Chemischen Gesellschaft*, 12(1):426–428, 1879. doi:10.1002/cber.187901201117.

F. M. Guerra, S. Bolotin, G. Lim, J. Heffernan, S. L. Deeks, Y. Li, and N. S. Crowcroft. The basic reproduction number (r_0) of measles: a systematic review. *The Lancet Infectious Diseases*, 17:e420–e428, 2017. doi:10.1016/s1473-3099(17)30307-9.

K. Gurung. *Fractal Dimension in Architecture: An Exploration of Spatial Dimension*. Master thesis, Anhalt University of Applied Sciences, Köthen, Germany, August 2017.

B. D. Hall and D. R. White. *An Introduction to Measurement Uncertainty*. Measurement Standards Laboratory of New Zealand, Lower Hutt, New Zealand, 2018. ISBN 978-0-473-40581-6. doi:10.5281/zenodo.3872590. URL `https://zenodo.org/record/3872590`. Also available from `www.lulu.com`.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, April 1970. doi:10.1093/biomet/57.1.97.

T. J. Heaton, M. Blaauw, P. G. Blackwell, C. Bronk Ramsey, P. J. Reimer, and E. M. Scott. The INTCAL20 approach to radiocarbon calibration curve construction: a new methodology using

Bayesian splines and errors-in-variables. *Radiocarbon*, pages 1–43, 2020. doi:10.1017/RDC.2020.46.

J. M. Heffernan, R. J. Smith, and L. M. Wahl. Perspectives on the basic reproductive ratio. *Journal of The Royal Society Interface*, 2: 281–293, 2005. doi:10.1098/rsif.2005.0042.

T. C. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386, November 2015. doi:10.1080/00031305.2015.1089789.

H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42:599–653, 2000. doi:10.1137/S0036144500371907.

J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Hoboken, NJ, second edition, 2019. ISBN 978-1-119-53662-8.

J. L. Hodges and E. L. Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34(2):598–611, June 1963. doi:10.1214/aoms/1177704172.

M. Hollander, D. A. Wolfe, and E. Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons, Hoboken, NJ, 3rd edition, 2014.

K. Hotokezaka et al. A Hubble constant measurement from superluminal motion of the jet in GW170817. *Nature Astronomy*, 3:940–944, July 2019. doi:10.1038/s41550-019-0820-1.

E. Hubble. A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15(3):168–173, 1929. doi:10.1073/pnas.15.3.168.

International Organization of Legal Metrology (OIML). *Weights of classes E_1, E_1, E_2, F_1, F_2, M_{1-2}, M_2, M_{2-3}, and M_3 — Part 1: Metrological and technical requirements*. Bureau International de Métrologie Légale (OIML), Paris, France, 2004. URL `https://www.oiml.org/en/files/pdf_r/r111-1-e04.pdf`. International Recommendation OIML R 111-1 Edition 2004 (E).

B. Ivanović, B. Milošević, and M. Obradović. *symmetry: Testing for Symmetry of Data and Model Residuals*, 2020. URL `https://CRAN.R-project.org/package=symmetry`. R package version 0.2.1.

H. Jeffreys. *Scientific Inference*. Cambridge University Press, London, third edition, 1973.

V. E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110:19313–19317, 2013. doi:10.1073/pnas.1313476110.

Joint Committee for Guides in Metrology. *Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008. URL 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 (JCGM). *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008. URL 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.

C. Kendall and T. B. Coplen. Distribution of oxygen-18 and deuterium in river waters across the United States. *Hydrological Processes*, 15:1363–1393, 2001. doi:10.1002/hyp.217.

W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115:700–721, August 1927. doi:10.1098/rspa.1927.0118.

D. V. Khmelev and F. J. Tweedie. Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16:299–307, 2001. doi:10.1093/llc/16.3.299.

C. Klein and B. Dutrow. *Manual of Mineral Science*. John Wiley & Sons, Hoboken, NJ, 23rd edition, 2007. ISBN 978-0-471-72157-4.

A. Koepke, T. Lafarge, A. Possolo, and B. Toman. Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia*, 54(3):S34–S62, 2017. doi:10.1088/1681-7575/aa6c0e.

T. Lafarge and A. Possolo. The NIST Uncertainty Machine. *NCSLI Measure Journal of Measurement Science*, 10(3):20–27, September 2015. doi:10.1080/19315775.2015.11721732.

D. Lara, J. Strickler, C. D. Olavarrieta, and C. Ellertson. Measuring induced abortion in Mexico: A comparison of four methodologies. *Sociological Methods & Research*, 32(4):529–558, May 2004. doi:10.1177/0049124103262685.

I. Lavagnini and F. Magno. A statistical overview on univariate calibration, inverse regression, and detection limits: Application to gas chromatography/mass spectrometry technique. *Mass Spectrometry Reviews*, 26(1):1–18, 2007. doi:10.1002/mas.20100.

G. Lemaître. Un univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques. *Annales de la Société Scientifique de Bruxelles A*, 47:49–59, 1927.

G. Lemaître. Republication of: A homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae. *General Relativity and Gravitation*, 45:1635–1646, 2013. doi:10.1007/s10714-013-1548-3.

A. Lindén and S. Mäntyniemi. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92 (7):1414–1421, 2011. doi:10.1890/10-1831.1.

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.

E. Lukacs. A characterization of the normal distribution. *Annals of Mathematical Statistics*, 13(1):91–93, March 1942. doi:10.1214/aoms/1177731647.

E. Macaulay et al. First cosmological results using Type Ia supernovae from the Dark Energy Survey: measurement of the Hubble constant. *Monthly Notices of the Royal Astronomical Society*, 486(2): 2184–2196, April 2019. doi:10.1093/mnras/stz978.

B. Mandelbrot. How long is the coast of Britain? statistical self-similarity and fractional dimension. *Science*, 156:636–638, May 1967. doi:10.1126/science.156.3775.636.

M. Martcheva. *An Introduction to Mathematical Epidemiology*, volume 61 of *Texts in Applied Mathematics*. Springer, New York, NY, 2010. ISBN 978-1-4899-7611-6. doi:10.1007/978-1-4899-7612-3.

J. A. Martin, B. E. Hamilton, M. J. K. Osterman, and A. K. Driscoll. Births: Final data for 2018. National Vital Statistics Reports 68(13), National Center for Health Statistics, Centers for Disease Control and Prevention (CDC), Hyattsville, MD, November 2019.

M.D. Mastrandrea, K.J. Mach, G. Plattner, O. Edenhofer, T. F. Stocker, C. B. Field, K. L. Ebi, and P.R. Matschoss. The IPCC AR5 guidance

note on consistent treatment of uncertainties: a common approach across the working groups. *Climatic Change*, 108:675–691, 2011. doi:10.1007/s10584-011-0178-6. Special Issue: Guidance for Characterizing and Communicating Uncertainty and Confidence in the Intergovernmental Panel on Climate Change.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953. doi:10.1063/1.1699114.

W. Miao, Y. R. Gel, and J. L. Gastwirth. A new test of symmetry about an unknown median. In A. C. Hsiung, Z. Ying, and C.-H. Zhang, editors, *Random Walk, Sequential Analysis and Related Topics: A Festschrift in Honor of Yuan-Shih Chow*, pages 199–214. World Scientific, Singapore, 2006. doi:10.1142/9789812772558_0013.

C. Michotte, G. Ratel, S. Courte, K. Kossert, O. Nähle, R. Dersch, T. Branger, C. Bobin, A. Yunoki, and Y. Sato. BIPM comparison BIPM.RI(II)-K1.Fe-59 of activity measurements of the radionuclide 59Fe for the PTB (Germany), LNE-LNHB (France) and the NMIJ (Japan), and the linked APMP.RI(II)-K2.Fe-59 comparison. *Metrologia*, 57(1A):06003, January 2020. doi:10.1088/0026-1394/57/1a/06003.

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, February 2014. doi:10.1016/j.envsoft.2013.10.010.

J. W. Munch. *Method 524.2. Measurement of Purgeable Organic Compounds in Water by Capillary Column Gas Chromatography/Mass Spectrometry*. National Exposure Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, OH, 1995. Revision 4.1.

M. G. Natrella. *Experimental Statistics*. National Bureau of Standards, Washington, D.C., 1963. National Bureau of Standards Handbook 91.

J. A. Nelder and R. Mead. A simplex algorithm for function minimization. *Computer Journal*, 7:308–313, 1965. doi:10.1093/comjnl/7.4.308.

R G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8): 857–872, 1998. doi:10.1002/(sici)1097-0258(19980430)17:8<857::aid-sim777>3.0.co;2-e.

A. O'Hagan. The Bayesian Approach to Statistics. In T. Rudas, editor, *Handbook of Probability: Theory and Applications*, chapter 6. Sage Publications, Thousand Oaks, CA, 2008. ISBN 978-1-4129-2714-7. doi:10.4135/9781452226620.n6.

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.

D. W. Pesce, J. A. Braatz, M. J. Reid, A. G. Riess, D. Scolnic, J. J. Condon, F. Gao, C. Henkel, C. M. V. Impellizzeri, C. Y. Kuo, and K. Y. Lo. The Megamaser cosmology project. XIII. combined hubble constant constraints. *The Astrophysical Journal*, 891(1):L1, February 2020. doi:10.3847/2041-8213/ab75f0.

P. E. Pontius and J. M. Cameron. *Realistic Uncertainties and the Mass Measurement Process — An Illustrated Review*. Number 103 in NBS Monograph Series. National Bureau of Standards, Washington, DC, 1967. URL `http://nvlpubs.nist.gov/nistpubs/Legacy/MONO/nbsmonograph103.pdf`.

A. Possolo. *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. National Institute of Standards and Technology, Gaithersburg, MD, 2015. doi:10.6028/NIST.TN.1900. NIST Technical Note 1900.

A. Possolo. *Evaluating, Expressing, and Propagating Measurement Uncertainty for NIST Reference Materials*. National Institute of Standards and Technology, Gaithersburg, MD, 2020. doi:10.6028/NIST.SP.260.202. NIST Special Publication 260-202.

A. Possolo and H. K. Iyer. Concepts and tools for the evaluation of measurement uncertainty. *Review of Scientific Instruments*, 88(1): 011301, 2017. doi:10.1063/1.4974274.

V. Poulin, T. L. Smith, T. Karwal, and M. Kamionkowski. Early dark energy can resolve the hubble tension. *Physical Review Letters*, 122: 221301, June 2019. doi:10.1103/PhysRevLett.122.221301.

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. doi:10.1016/j.dental.2009.09.006.

C. Bronk Ramsey. Bayesian analysis of radiocarbon dates. *Radiocarbon*, 51(1):337–360, 2009. doi:10.1017/S0033822200033865.

M. J. Reid, D. W. Pesce, and A. G. Riess. An improved distance to NGC 4258 and its implications for the Hubble Constant. *The Astrophysical Journal*, 886(2):L27, November 2019. doi:10.3847/2041-8213/ab552d.

P. J. Reimer and et al. The INTCAL20 northern hemisphere radiocarbon age calibration curve (0–55 cal kbp). *Radiocarbon*, pages 1–33, 2020. doi:10.1017/RDC.2020.41.

A. G. Riess, S. Casertano, W. Yuan, L. M. Macri, and D. Scolnic. Large Magellanic Cloud Cepheid Standards provide a 1% foundation for the determination of the Hubble Constant and stronger evidence for physics beyond λCDM. *The Astrophysical Journal*, 876(1):85, May 2019. doi:10.3847/1538-4357/ab1422.

C. Robert and G. Casella. A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26(1):102–115, 2011. doi:10.1214/10-STS3510.

J. Ryan, Y. Chen, and B. Ratra. Baryon acoustic oscillation, Hubble parameter, and angular size measurement constraints on the Hubble constant, dark energy dynamics, and spatial curvature. *Monthly Notices of the Royal Astronomical Society*, 488(3):3844–3856, July 2019. doi:10.1093/mnras/stz1966.

L. J. Savage. *The Foundations of Statistics*. Dover Publications, New York, New York, 1972.

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

G. A. F. Seber. *A Matrix Handbook for Statisticians*. John Wiley & Sons, Hoboken, NJ, 2008. ISBN 978-0-471-74869-4.

A. J. Shajib et al. STRIDES: A 3.9 per cent measurement of the hubble constant from the strong lens system DES J0408-5354, 2019. URL `https://arxiv.org/abs/1910.06306`.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3,4):591–611, 1965. doi:10.2307/2333709.

R. Silberzahn and E. L. Uhlmann. Crowdsourced research: Many hands make tight work. *Nature*, 526:189–191, October 2015. doi:10.1038/526189a.

J. Silver-Greenberg and S. Cowley. 5 Reasons to Question Breath Tests. *The New York Times*, November 3, 2019. Business Section.

G. Simpson. Accuracy and precision of breath-alcohol measurements for a random subject in the postabsorptive state. *Clinical Chemistry*, 33(2):261–268, 1987. doi:10.1093/clinchem/33.2.261.

T. Simpson. A Letter to the Right Honourable George Earl of Macclesfield, President of the Royal Society, on the Advantage of Taking the Mean of a Number of Observations, in Practical Astronomy. *Philosophical Transactions of the Royal Society of London*, 49:82–93, 1755. doi:10.1098/rstl.1755.0020.

J. Sokol. A recharged debate over the speed of the expansion of the universe could lead to new physics. *Science*, March 2017. doi:10.1126/science.aal0877.

T. Stockman, G. Monroe, and S. Cordner. Venus is not earth's closest neighbor. *Physics Today*, 72, March 2019. doi:10.1063/PT.6.3.20190312a.

L. D. Stone, C. Keller, T. L. Kratzke, and J. Strumpfer. Search analysis for the location of the AF447 underwater wreckage. Technical report, Metron Scientific Solutions, Reston, VA, January 2011. Report to Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile.

Stan Development Team. *Stan User's Guide*. `mc-stan.org`, 2019. Stan Version 2.23.

Teva Pharmaceuticals USA, Inc. v. *Sandoz, Inc.* 574 U. S. 318 (2015), 2015.

W. L. Tew and G. F. Strouse. *Standard Reference Material 1750: Standard Platinum Resistance Thermometers, 13.8033 K to 429.7485 K*. NIST Special Publication 260-139. National Institute of Standards and Technology, Gaithersburg, MD, November 2001. doi:10.6028/NIST.SP.260-139.

M. Thompson and S. L. R. Ellison. Dark uncertainty. *Accreditation and Quality Assurance*, 16:483–487, October 2011. doi:10.1007/s00769-011-0803-0.

H. L. Thuillier and R. Smyth. *A Manual of Surveying for India, detailing the mode of operations on the Trigonometrical, Topographical, and Revenue Surveys of India*. Thacker, Spink & Co., Calcutta, India, third edition, 1875.

J. Todd. The prehistory and early history of computation at the U.S. National Bureau of Standards. In S. G. Nash, editor, *A History of Scientific Computing*, ACM Press History Series, pages 251–268. Addison-Wesley, Reading, MA, 1990. Conference on the History of Scientific and Numeric Computation, Princeton, N.J., 1987.

J. R. Townsley. BP: Time for a change. *Radiocarbon*, 59(1):177–178, 2017. doi:10.1017/RDC.2017.2.

V. Trimble. H_0: The incredible shrinking constant, 1925-1975. *Publications of the Astronomical Society of the Pacific*, 108:1073–1082, December 1996. doi:10.1086/133837.

J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.

J. W. Tukey. Choosing techniques for the analysis of data. In L. V. Jones, editor, *The Collected Works of John Tukey — Philosophy and Principles of Data Analysis: 1965-1986*, volume 4, chapter 24. Wadsworth & Brooks/Cole, Monterey, CA, 1986. Previously unpublished manuscript.

R. N. Varner and R. C. Raybold. *National Bureau of Standards Mass Calibration Computer Software*. NIST Technical Note 1127. National Bureau of Standards, Washington, DC, July 1980. URL https://nvlpubs.nist.gov/nistpubs/Legacy/TN/nbstechnicalnote1127.pdf.

A. Vehtari, J. Gabry, Y. Yao, and A. Gelman. loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models, 2019. URL https://CRAN.R-project.org/package=loo. R package version 2.2.0.

G. H. White. Basics of estimating measurement uncertainty. *The Clinical Biochemist Reviews*, 29 (Supplement 1):S53–S60, August 2008.

G. H. White, C. A. Campbell, and A. R Horvath. Is this a Critical, Panic, Alarm, Urgent, or Markedly Abnormal Result? *Clinical Chemistry*, 60(12):1569–1570, December 2014. doi:10.1373/clinchem.2014.227645.

WHO. Preventing unsafe abortion. Evidence Brief WHO/RHR/19.21, World Health Organization, Geneva, Switzerland, 2019.

E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212, 1927. doi:10.2307/2276774.

K. C. Wong et al. H0LiCOW XIII. A 2.4 % measurement of h_0 from lensed quasars: 5.3σ tension between early and late-Universe probes. arXiv:1907.04869, 2019. URL https://arxiv.org/abs/1907.04869.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*, 2017. doi:10.1214/17-BA1091.

Y. Ye. *Interior Point Algorithms: Theory and Analysis*. John Wiley & Sons, New York, NY, 1997. ISBN 978-0471174202.

H. Zangl, M. Zine-Zine, and K. Hoermaier. Utilization of software tools for uncertainty calculation in measurement science education. *Journal of Physics: Conference Series*, 588:012054, 2015. doi:10.1088/1742-6596/588/1/012054.

M. Zelen. Linear estimation and related topics. In J. Todd, editor, *Survey of Numerical Analysis*, chapter 17, pages 558–584. McGraw-Hill, New York, NY, 1962.