ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Bailey DC. 2017 Not Normal: the uncertainties of scientific measurements. *R. Soc. open sci.* **4**: 160600. http://dx.doi.org/10.1098/rsos.160600

Received: 18 August 2016
Accepted: 30 November 2016

Subject Category:
Mathematics

Subject Areas:
statistics/complexity

Keywords:
measurement uncertainty, research reproducibility, systematic errors, complex systems, meta-analysis, metrology

Author for correspondence:
David C. Bailey
e-mail: dbailey@physics.utoronto.ca

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3660959.

THE ROYAL SOCIETY
PUBLISHING

Not Normal: the uncertainties of scientific measurements

David C. Bailey

Department of Physics, University of Toronto, Toronto, Ontario, Canada M5S 1A7

DCB, 0000-0002-7970-7839

Judging the significance and reproducibility of quantitative research requires a good understanding of relevant uncertainties, but it is often unclear how well these have been evaluated and what they imply. Reported scientific uncertainties were studied by analysing 41 000 measurements of 3200 quantities from medicine, nuclear and particle physics, and interlaboratory comparisons ranging from chemistry to toxicology. Outliers are common, with 5σ disagreements up to five orders of magnitude more frequent than naively expected. Uncertainty-normalized differences between multiple measurements of the same quantity are consistent with heavy-tailed Student's *t*-distributions that are often almost Cauchy, far from a Gaussian Normal bell curve. Medical research uncertainties are generally as well evaluated as those in physics, but physics uncertainty improves more rapidly, making feasible simple significance criteria such as the 5σ discovery convention in particle physics. Contributions to measurement uncertainty from mistakes and unknown problems are not completely unpredictable. Such errors appear to have power-law distributions consistent with how designed complex systems fail, and how unknown systematic errors are constrained by researchers. This better understanding may help improve analysis and meta-analysis of data, and help scientists and the public have more realistic expectations of what scientific results imply.

1. Introduction

What do reported uncertainties actually tell us about the accuracy of scientific measurements and the likelihood that different measurements will disagree? No scientist expects different research studies to always agree, but the frequent failure of published research to be confirmed has generated much concern about scientific reproducibility [1,2].

When scientists investigate many quantities in very large amounts of data, interesting but ultimately false results may occur by chance and are often published. In particle physics, bitter experience with frequent failures to confirm such results eventually led to an ad hoc ‘5-sigma’ discovery criterion [3–6], i.e. a ‘discovery’ is only taken seriously if the estimated probability for observing the result without new physics is less than the chance of a single sample from a Normal distribution being more than five standard deviations (5σ) from the mean.

In other fields, arguments that most novel discoveries are false [7] have caused increased emphasis on reporting the value and uncertainty of measured quantities, not just whether the value is statistically different from zero [8,9]. Research confirmation is then judged by how well independent studies agree according to their reported uncertainties, so assessing reproducibility requires accurate evaluation and realistic understanding of these uncertainties. This understanding is also required when analysing data, combining studies in meta-analyses, or making scientific, business or policy judgements based on research. The experience of research fields like physics, in which values and uncertainties have long been regularly reported, may provide some guidance on what reproducibility can reasonably be expected [10].

Most recent investigations into reproducibility focus on how often observed effects disappear in subsequent research, revealing strong selection bias in published results. Removing such bias is extremely important, but may not reduce the absolute number of false discoveries since not publishing non-significant results does not make the ‘discoveries’ go away. Controlling the rate of false discoveries depends on establishing criteria that reflect real measurement uncertainties, especially the likelihood of extreme fluctuations and outliers [11].

Outliers are observations that disagree by an abnormal amount with other measurements of the same quantity. Despite every scientist knowing that the rate of outliers is always greater than naively expected, there is no widely accepted heuristic for estimating the size or shape of these long tails. These estimates are often assumed to be approximately Normal (Gaussian), but it is easy to find examples where this is clearly untrue [12–14].

To examine the accuracy of reported uncertainties, this paper reviews multiple published measurements of many different quantities, looking at the differences between measurements of each quantity normalized by their reported uncertainties. Previous similar studies [15–20] reported on only a few hundred to a few thousand measurements, mostly in subatomic physics. This study reports on how well multiple measurements of the same quantity agree, and hence what are reasonable expectations for the reproducibility of published scientific measurements. Of particular interest is the frequency of large disagreements which usually reflect unexpected systematic effects.

1.1. Systematic effects

Sources of uncertainty are often categorized as statistical or systematic, and their methods of evaluation classified as Type A or B [21]. Type A evaluations are based on observed frequency distributions; whereas Type B evaluations use other methods. Statistical uncertainties are always evaluated from primary data using Type A methods, and can, in principle, be made arbitrarily small by repeated measurement or large enough sample size.

Uncertainties due to systematic effects may be evaluated by either Type A or B methods, and fall into several overlapping classes [22]. Class 1 systematics, which include many calibration and background uncertainties, are evaluated by Type A methods using ancillary data. Class 2 systematics are almost everything else that might bias a measurement, and are caused by a lack of knowledge or uncertainty in the measurement model, such as the reading error of an instrument or the uncertainties in Monte Carlo estimates of corrections to the measurement. Class 3 systematics are theoretical uncertainties in the interpretation of a measurement. For example, determining the proton radius using the Lamb shift of muonic hydrogen requires over 20 theoretical corrections [23] that are potential sources of uncertainty in the proton radius, even if the actual measurement of the Lamb shift is perfect. The uncertainties associated with Classes 2 and 3 systematic effects cannot be made arbitrarily small by simply getting more data.

When considering the likelihood of extreme fluctuations in measurements, mistakes and ‘unknown unknowns’ are particularly important, but they are usually assumed to be statistically intractable and are not often considered in traditional uncertainty analysis. Mistakes are ‘unknown knowns’, i.e. something that is thought to be known but is not, and it is believed that good scientists should not make mistakes.

‘Unknown unknowns’ are factors that affect a measurement but are unknown and unanticipated based on past experience and knowledge [24]. For example, during the first 5 years of operation of LEP (the Large Electron Positron collider), the effect of local railway traffic on measurements of the Z^0 boson mass was an ‘unknown unknown’ that no-one thought about. Then improved monitoring revealed unexpected variations in the accelerator magnetic field, and after much investigation these variations were found to be caused by electric rail line ground leakage currents flowing through the LEP vacuum pipe [25].

In general, systematic effects are challenging to estimate [12,21,22,26–29], but can be partially constrained by researchers making multiple internal and external consistency checks: is the result compatible with previous data or theoretical expectations? Is the same result obtained for different times, places, assumptions, instruments or subgroups? As described by Dorsey [30, p. 11], scientists 'change every condition that seems by any chance likely to affect the result, and some that do not, in every case pushing the change well beyond any that seems at all likely'. If an inconsistency is observed and its cause understood, the problem can often be fixed and new data taken, or the effect monitored and corrections made. If the cause cannot be identified, however, then the observed dispersion of values must be included in the uncertainty.

The existence of unknown systematic effects or mistakes may be revealed by consistency checks [31], but small unknown systematics and mistakes are unlikely to be noticed if they do not affect the measurement by more than the expected uncertainty. Even large problems can be missed by chance (see §4.4) or if the conditions changed between consistency checks do not alter the size of the systematic effect. The power of consistency checks is limited by the impossibility of completely changing all apparatus, methods, theory and researchers between measurements, so one can never be certain that all significant systematic effects have been identified.

2. Methods

2.1. Data

Quantities were only included in this study if they are significant enough to have warranted at least five independent measurements with clearly stated uncertainties.

Medical and health research data were extracted from some of the many meta-analyses published by the Cochrane Collaboration [32]; a total of 5580 measurements of 310 quantities generating 99 433 comparison pairs were included. Particle physics data (8469 measurements, 864 quantities and 53 988 pairs) were retrieved from the Review of Particle Physics [33,34]. Nuclear physics data (12 380 measurements, 1437 quantities and 66 677 pairs) were obtained from the Table of Radionuclides [35].

Most nuclear and particle physics measurements have prior experimental or theoretical expectations which may influence results from nominally independent experiments [20,36], and medical research has similar biases [7], so this study also includes a large sample of interlaboratory studies that do not have precise prior expectations for their results. In these studies, multiple independent laboratories measure the same quantity and compare results. For example, the same mass standard might be measured by national laboratories in different countries, or an unknown archaeological sample might be divided and distributed to many laboratories, with each laboratory reporting back its Carbon-14 measurement of the sample's age. None of the laboratories knows the expected value for the quantity nor the results from other laboratories, so there should be no expectation, selection or publication biases. These Interlab studies (14 097 measurements, 617 quantities and 965 416 pairs) were selected from a wide range of sources in fields such as analytical chemistry, environmental sciences, metrology and toxicology. The measurements ranged from genetic contamination of food to high-precision comparison of fundamental physical standards, and were carried out by a mix of national, university and commercial laboratories.

All quantities analysed are listed in the electronic supplementary materials [37].

2.2. Data selection and collection

Data were entered using a variety of semi-automatic scripts, optical-character recognition and manual methods. No attempt was made to recalculate past results based on current knowledge, or to remove results that were later retracted or amended, since the original paper was the best result at the time it was published. When the Review of Particle Physics [34] noted that earlier data had been dropped, the missing results were retrieved from previous editions [38].

To ensure that measurements were as independent as possible, measurements were excluded if they were obviously not independent of other data already included. Because relationships between measurements are often obscure, however, there undoubtedly remain many correlations between the published results used.

Medical and health data were selected from the 8105 reviews in the Cochrane database [32] as of 25 September 2013. Data were analysed from 221 Intervention Reviews whose abstract mentioned at least six trials with more than 5000 total participants, and which had at least one analysis with five or more studies with greater than 3500 total participants. The average heterogeneity inconsistency index ($I^2 \equiv 1 - \text{d.f.}/\chi^2$) [36,39] is about 40% for the analyses reported here. Because analyses within a review may be correlated, only a maximum of three analyses and five comparison groups were included from any one review. About 80% of the Cochrane results are the ratio of intervention and control binomial probabilities, e.g. mortality rates for a drug and a placebo. Such ratios are not Normal [40], so they were converted to differences that should be Normal in the Gaussian limit, i.e. when the group size n and probability p are such that n, np and $(1-p)n$ are all $\gg 1$, so the binomial distribution converges towards a Gaussian distribution. (The median observed values for these data were $n = 100$, $p = 0.16$.) The 68.3% binomial probability confidence interval was calculated for both the intervention and control groups to determine the uncertainties.

2.3. Uncertainty evaluation

Measurements with uncertainties are typically reported as $x \pm u$, which means that the interval $x - u$ to $x + u$ contains with some defined probability 'the values that could reasonably be attributed to the measurand' [21]. Most frequently, uncertainty intervals are given as $\pm k u_S$, where k is the coverage factor and u_S is the 'standard uncertainty', i.e. the uncertainty of a measurement expressed as the standard deviation of the expected dispersion of values. Uncertainties in particle physics and medicine are often instead reported as the bounds of either 68.3% or 95% confidence intervals, which for a Normal distribution are equivalent to the $k = 1$ and 2 standard uncertainty intervals.

For this study, all uncertainties were converted to nominal 68.3% confidence interval uncertainties. The vast majority of measurements reported simple single uncertainties, but if more than a single uncertainty was reported, e.g. 'statistical' and 'systematic', they were added in quadrature.

2.4. Normalized differences

All measurements, $x_i \pm u_i$, of a given quantity were combined in all possible pairs and the difference between the two measurements of each pair calculated in units of their combined uncertainty u_{ij}:

$$z_{ij} = \frac{|x_i - x_j|}{\sqrt{u_i^2 + u_j^2}}. \tag{2.1}$$

The dispersion of z_{ij} values can be used to judge whether independent measurements of a quantity are 'compatible' [41]. A feature of z as a metric for measurement agreement is that it does not require a reference value for the quantity. (The challenges and effects of using reference values are discussed in §3.3.)

The uncertainties in equation (2.1) are combined in quadrature, as expected for standard uncertainties of independent measurements. (The effects of any lack of independence are discussed in §3.2.)

Uncertainties based on confidence intervals may not be symmetric about the reported value, which is the case for about 13% of particle, 6% of medical, 0.3% of nuclear and 0.06% of interlab measurements. Following common (albeit imperfect) practice [42], if the reported plus and minus uncertainties were asymmetric, z_{ij} was calculated from equation (2.1) using the uncertainty for the side towards the other member of the comparison pair. For example, if $x_1 = 80\pm^{3}_{2}$, $x_2 = 100\pm^{5}_{4}$ and $x_3 = 126\pm^{15}_{12}$, then $z_{12} = (100 - 80)/\sqrt{3^2 + 4^2}$ and $z_{23} = (126 - 100)/\sqrt{5^2 + 12^2}$.

The distributions of the z_{ij} differences are histogrammed in figure 1, with each pair weighted such that the total weight for a quantity is the number of measurements of that quantity. For example, if a quantity has 10 measurements, there are 45 possible pairs, and each entry has a weight of 10/45. (Other weighting schemes are discussed in §3.3.) The final frequency distribution within each research area is then normalized so that its total observed probability adds up to 1. If the measurement uncertainties are well evaluated and correspond to Normally distributed probabilities for x, then z is expected to be Normally distributed with a standard deviation $\sigma = 1$.

Probability distribution uncertainties (e.g. the vertical error bars in figure 1) were evaluated using a bootstrap Monte Carlo method where quantities were drawn randomly with replacement from the actual dataset until the number of Monte Carlo quantities equalled the actual number of quantities. The resulting artificial dataset was histogrammed, the process repeated 1000 times, and the standard deviations of the Monte Carlo probabilities calculated for each z bin.

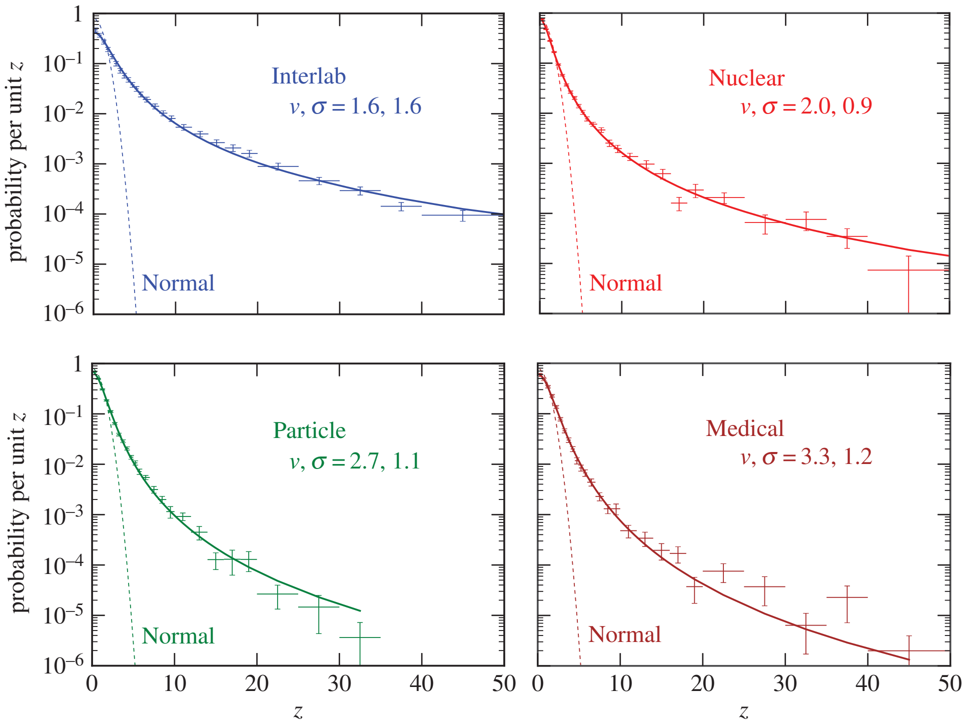

Figure 1. Histograms of uncertainty normalized differences (z_{ij} from equation (2.1)) per unit of z. Horizontal and vertical error bars are the bin width and the standard uncertainty evaluated by a bootstrap Monte Carlo. The smooth curves are best-fit Student's t-distributions. The dashed curves are Normal distributions.

Random selection of measurements instead of quantities was not chosen for uncertainty evaluation because of the corrections then required to avoid bias and artefacts. For example, if measurements are randomly drawn, a quantity with only five measurements will often be missing from the artificial dataset for having too few (less than five) measurements drawn, or if it does have five measurements some of them will be duplicates generating unrealistic $z = 0$ values, or if duplicates are excluded then they will always be the same five non-random measurements. Without correcting for such effects, the resulting measurement Monte Carlo generated uncertainties are too small to be consistent with the observed bin-to-bin fluctuations in figure 1. Correcting for such effects requires using characteristics of the actual quantities and would be effectively equivalent to using random quantities.

2.5. Data fits

Attempts were made to fit the data to a wide variety of functions, but by far the best fits were to non-standardized Student's t-probability density distributions with ν degrees of freedom.

$$S_{\nu,\sigma}(z) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}\sigma} \frac{1}{(1+(z/\sigma)^2/\nu)^{(\nu+1)/2}}. \tag{2.2}$$

Student's t-distribution is essentially a smoothly symmetric normalizable power law, with $S_{\nu,\sigma}(z) \sim (z/\sigma)^{-(\nu+1)}$ for $|z| \gg \sigma\sqrt{\nu}$.

The fitted parameter σ defines the core width and overall scale of the distribution and is equal to the standard deviation in the $\nu \to \infty$ Gaussian limit and to the half-width at half maximum in the $\nu \to 1$ Cauchy (also known as Lorentzian or Breit-Wigner) limit. The parameter ν determines the size of the tails, with small ν corresponding to large tails. The values and standard uncertainties in σ and ν were determined from a nonlinear least-squares fit to the data that minimizes the nominal χ^2 [43]:

$$\chi^2 = \sum_{i=1}^{N_{\text{bins}}} \frac{(B_i - S_{\nu,\sigma}(z_i))^2}{u_{B_i}^2}, \tag{2.3}$$

where z_i, B_i and u_{B_i} are the bin z, contents, and uncertainties of the observed z-distributions shown in figure 1.

Possible values of z are sometimes limited by the allowed range of measurement values, which could suppress heavy tails. For example, many quantities are fractions that must lie between 0 and 1, and there is less room for two measurements with 10% uncertainty to disagree by 5σ than for two 0.01% measurements. The size of this effect was estimated using Monte Carlo methods to generate simulated data based on the values and uncertainties of the actual data, constrained by any obvious bounds on their allowed values. The simulated data were then fit to see if applying the bounds changed the fitted values for σ and ν. The largest effect was for Medical data where ν was reduced by about 0.1 when minimally restrictive bounds were assumed. Stronger bounds might exist for some quantities, but determining them would require careful measurement-by-measurement assessment beyond the scope of this study. For example, each measurement of the long-term duration of the effect of a medical drug or treatment would have an upper bound set by the length of that study. Since correcting for bounds can only make ν smaller (corresponding to even heavier tails), and the observed effects were negligible, no corrections were applied to the values of ν reported here.

3. Results

3.1. Observed distributions

Histograms of the z-distributions for different datasets are shown in figure 1. The complementary cumulative distributions of the data are given in table 1 and shown in figure 2.

None of the data are close to Gaussian, but all can reasonably be described by almost-Cauchy Student's t-distributions with $\nu \sim 2-3$. For comparison, fits to these data with Lévy stable distributions have nominal χ^2 4–30 times worse than the fits to Student's t-distributions. The number of '5σ' (i.e. $z > 5$) disagreements observed is as high as 0.12, compared with the 6×10^{-7} expected for a Normal distribution.

The fitted values for ν and σ are shown in table 2. Also shown in table 2 are two data subsets expected to be of higher quality, BIPM Interlaboratory Key comparisons (372 quantities, 3712 measurements and 20 245 pairs) and Stable Particle properties (335 quantities, 3041 measurements and 16 649 pairs). The Key comparisons [44] should define state-of-the-art accuracy, since they are measurements of important metrological standards carried out by national laboratories. Stable particles are often easier to study than other particles, so their properties are expected to be better determined. Both 'better' data subsets do have narrower distributions consistent with higher quality, but they still have heavy tails. More selected data subsets are discussed in §3.4.

The probability distribution for the nominal χ^2 statistic is not expected to be an exact regular χ^2-distribution. The differences are due to the non-Gaussian uncertainties of the low-population high-z bins, and because the bin contents are not independent since a single measurement can contribute to multiple bins as part of different permutation pairs. Based on fits of simulated datasets with a mix of ν comparable to the observed data, the range of nominal χ^2 reported in table 2 seems reasonable, i.e. the chances of $\chi^2/\mathrm{d.f.} \leq 0.6$ or ≥ 1.9 were 15% and 2%, respectively.

To see whether more important quantities are measured with less disagreement, a small additional dataset of measurements of fundamental physical constants (7 quantities, 320 measurements and 9098 pairs) was also analysed. The constants are Avogadro's number, the fine structure constant, the Planck constant, Newton's gravitational constant, the deuteron binding energy, the Rydberg constant and the speed of light (before it became a defined constant). These measurements have very heavy tails, despite their importance in physical science. Quantities with more interest do not seem to be better measured, as is also shown by considering only quantities with at least 10 published measurements, which do not have significantly smaller tails (see σ_{10},ν_{10} in table 2).

Figure 1 shows that the comparison pairs z_{ij} are Student's t-distributed, but what does this imply about the dispersion of individual x_i measurements? Except for the $\nu = 1$ and ∞ Cauchy and Normal limits, the distribution of differences of values selected from a Student's t-distribution is not itself a t-distribution, but it can be closely approximated as one [45]. The distributions of the parent individual x measurements were estimated by Monte Carlo deconvolution. Artificial measurements were generated from t-distributions with parameters ν_x and σ_x, and these measurements combined into permutation pairs to generate an artificial z-distribution. This distribution was compared to the observed z-distributions, and then ν_x and σ_x were iteratively adjusted until the best match was achieved between the artificial and observed z-distributions. As shown in table 2, the approximate Student's t-parameters

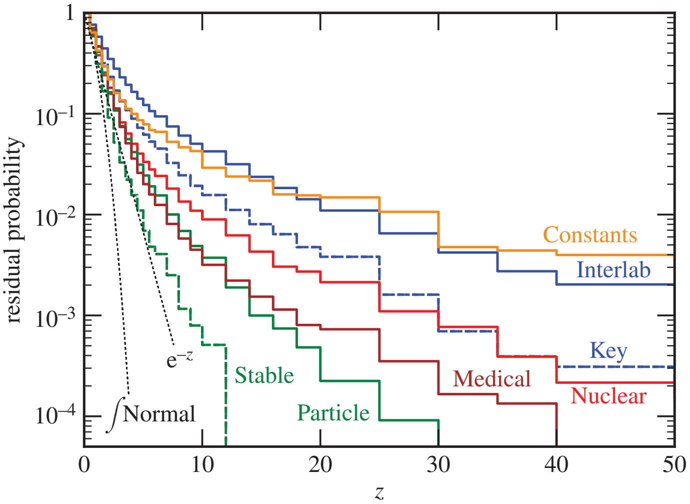

Figure 2. The observed probability of two measurements disagreeing by more than z standard uncertainties for different datasets: $\int_z^\infty \mathcal{P}(x)\,\mathrm{d}x$ (see also table 1).

Table 1. Observed chance of experimental disagreement by more than z standard uncertainties for different datasets, compared to values expected for some theoretical distributions. Also listed are the z-values that bound 95% of the distribution, i.e. $p_{\text{true}} = 0.05$, and p-values for that z for a Normal distribution.

$z >$	1	2	3	5	10	$z_{0.95}$	$p_{\text{Normal}}(z_{0.95})$
Interlab	0.58	0.35	0.23	0.12	0.042	9.0	2×10^{-19}
(key)	0.46	0.23	0.13	0.062	0.016	5.7	1×10^{-8}
Nuclear	0.38	0.16	0.082	0.033	0.009	4.0	6×10^{-5}
Particle	0.41	0.16	0.075	0.024	0.004	3.7	2×10^{-4}
(Stable)	0.31	0.091	0.033	0.007	0.0005	2.5	1×10^{-2}
Medical	0.47	0.18	0.074	0.020	0.003	3.5	4×10^{-4}
Constants	0.42	0.22	0.14	0.078	0.029	7.2	6×10^{-13}
Normal (Gaussian)	0.32	0.046	0.0027	5.7×10^{-7}	1.5×10^{-23}	1.96	5×10^{-2}
Student's t ($\nu = 10$)	0.34	0.073	0.013	5.4×10^{-4}	1.6×10^{-6}	2.23	2.6×10^{-2}
exponential	0.37	0.14	0.050	0.007	4.5×10^{-5}	3.0	2.7×10^{-3}
Student's t ($\nu = 2$)	0.42	0.18	0.095	0.038	0.010	4.3	2×10^{-15}
Cauchy	0.50	0.30	0.20	0.13	0.063	12.8	2×10^{-37}

(ν_x, σ_x) of the individual measurement populations have $\nu_x < \nu$ and hence are slightly more Cauchy-like than the permutation pairs distributions.

3.2. Combined uncertainty

The definition of z by equation (2.1) assumes that the measurements x_i and x_j are independent and that the uncertainties u_i and u_j can be combined following the rules for standard uncertainties.

If x_i and x_j are correlated, however, equation (2.1) should be replaced by

$$z_{ij} = \frac{|x_i - x_j|}{\sqrt{u_i^2 - 2\mathrm{cov}(x_i, x_j) + u_j^2}}, \tag{3.1}$$

where $\mathrm{cov}(x_i, x_j)$ is the covariance of x_i and x_j [21].

Table 2. Fitted Student's t-parameters with nominal χ^2 per degree of freedom. Also shown are parameters for quantities with greater than or equal to 10 measurements, for newer measurements made since the year 2000, and for the approximate distribution of individual measurements. Uncertainties not shown for σ_{10}, σ_{new}, ν_x and σ_x are $\lesssim$0.1. Asterisk indicates value if Cauchy uncertainties are added in quadrature.

	ν	σ	χ^2/d.f.	ν_{10}	σ_{10}	ν_{new}	σ_{new}	ν_x	σ_x
Interlab	1.64 ± 0.05	1.62 ± 0.05	1.1	1.6 ± 0.1	1.7	1.6 ± 0.1	1.7	1.5	1.3
Key	1.90 ± 0.10	1.12 ± 0.04	1.9	1.7 ± 0.1	1.1	1.9 ± 0.1	1.2	1.7	0.9
Nuclear	1.99 ± 0.06	0.90 ± 0.02	1.6	2.4 ± 0.2	1.1	2.1 ± 0.2	0.9	1.8	0.7
Particle	2.75 ± 0.10	1.05 ± 0.02	1.5	2.8 ± 0.1	1.1	2.6 ± 0.2	1.0	2.4	0.9
Stable	3.45 ± 0.16	0.86 ± 0.02	0.6	3.8 ± 0.4	0.9	7.6 ± 1.3	0.9	2.9	0.8
Medical	3.30 ± 0.11	1.18 ± 0.02	0.7	3.3 ± 0.1	1.2	3.2 ± 0.2	1.2	2.8	1.0
Constants	1.81 ± 0.15	0.89 ± 0.06	0.8	1.8 ± 0.2	0.9	1.3 ± 0.3	1.1	1.7	0.7
Normal	∞	1.0						∞	1.0
Cauchy	1.0	$\sqrt{2}^*$						1.0	1.0

It is not in general possible to quantitatively evaluate the covariance for individual pairs of measurements in the datasets, but the effects of any correlations are not expected to be large, and they cannot explain the observed heavy tails. Any positive covariance would decrease the denominator in equation (3.1) and increase the width of the z-distributions. Correlations between measurements are expected to be much more likely positive than negative, but even perfect anti-correlation could only decrease z-values by at most a factor of $1/\sqrt{2}$ compared to the uncorrelated case (i.e. changing $\text{cov}(x_i, x_j)$ from 0 to $-u_i u_j$ in equation (3.1) reduces z_{ij} by $\sqrt{2}$ if $u_i = u_j$, and less if $u_i \neq u_j$). Correlations are further discussed in §3.6.

Another possible issue with equation (2.1) is that its usual derivation assumes that u_i and u_j are standard deviations of the expected dispersion of possible values (e.g. [21, §E.3.1]). This assumption is a concern since the standard deviation is an undefined quantity for Student's t-distributions if $\nu < 2$, and the observed z and inferred x-distributions have ν near or below this value. Even if the variance of a distribution is undefined, however, the dispersion of the difference of two independent variables drawn from such distributions may still be calculated numerically and in some cases analytically.

Cauchy uncertainties add linearly instead of in quadrature, since the distribution of differences of two variables drawn from two Cauchy distributions with widths σ_1 and σ_2 is simply another Cauchy distribution with width $\sigma_{\text{diff}} = \sigma_1 + \sigma_2$. The corresponding definition of z would be

$$z_{ij}^{\text{Cauchy}} = \frac{|x_i - x_j|}{u_i + u_j}. \tag{3.2}$$

Almost-Cauchy distributions should almost follow the rules for combination of Cauchy ($\nu = 1$) distributions. Applying equation (3.2) to the data produces z-distributions that appear almost identical to those in figure 1, except that the fitted values of σ for Interlab, nuclear, particle, medical data are smaller by factors of $0.78, 0.80, 0.75, 0.74$, while the the fitted values of ν are almost unchanged ($\nu_{\text{linear}}/\nu_{\text{quad}} = 0.99, 1.00, 0.98, 0.94$). The scale factor for σ would be $1/\sqrt{2} = 0.71$ if all measurements of a quantity had equal uncertainties ($u_i = u_j$), since switching from quadrature (equation (2.1)) to linear (equation (3.2)) would simply scale all the calculated z-values by $1/\sqrt{2}$ and not affect ν. Similarly, if data with equal $\nu = 1, \sigma = 1$ Cauchy uncertainties were analysed using equation (2.1), the resulting permutation pairs would have $\nu = 1, \sigma = \sqrt{2}$, as shown in the last line of table 2.

3.3. Alternative weighting schemes and compatibility measures

There are several ways to weight data in the distribution plots, but the fitted parameter values are not usually greatly affected by the choice (table 3). The default method was to give each measurement equal weight ('M' in table 3). Jeng [20] gave equal weight to all measurement pairs ('P'), but this gives extreme weight to quantities with a large number (N) of measurements since the number of permutations grows

Table 3. Fitted Student's *t*-parameters for weighting by quantities (Q), measurements (M, the default), permutations (P) or using difference from weighted mean (h).

	ν	σ	χ^2/d.f.
Interlab			
Q	1.65 ± 0.12	1.35 ± 0.08	4.0
M	1.64 ± 0.05	1.62 ± 0.05	1.1
P	1.70 ± 0.04	1.76 ± 0.05	0.3
h	1.09 ± 0.08	2.06 ± 0.13	1.3
Nuclear			
Q	1.93 ± 0.07	0.85 ± 0.02	1.8
M	1.99 ± 0.06	0.90 ± 0.02	1.6
P	2.19 ± 0.07	0.98 ± 0.03	1.4
h	1.82 ± 0.06	0.95 ± 0.02	1.4
Particle			
Q	2.76 ± 0.11	1.01 ± 0.02	1.6
M	2.75 ± 0.10	1.05 ± 0.02	1.5
P	2.91 ± 0.09	1.14 ± 0.02	1.0
h	2.26 ± 0.12	1.10 ± 0.03	1.5
Medical			
Q	3.44 ± 0.16	1.24 ± 0.02	1.2
M	3.30 ± 0.11	1.18 ± 0.02	0.7
P	3.59 ± 0.12	1.17 ± 0.03	0.4
h	3.00 ± 0.18	1.21 ± 0.04	0.8

as $(N-1)N/2$. Giving each quantity equal weight ('Q') also seems less fair, since a quantity measured many times will be weighted the same as a quantity measured only a few times.

Instead of using measurement pairs to study compatibility, Roos *et al.* [16] instead calculated the weighted mean for each quantity, and then plotted the distribution of the uncertainty-normalized difference ('h') from that mean for each measurement, i.e.

$$h_i = \frac{|x_i - \bar{x}|}{\sqrt{u_i^2 + u_{\bar{x}}^2}}, \tag{3.3}$$

where

$$\bar{x} = \frac{\sum_i (x_i/u_i)^2}{\sum_i 1/u_i^2} \quad \text{and} \quad \frac{1}{u_{\bar{x}}^2} = \sum_i \frac{1}{u_i^2} \tag{3.4}$$

h is very similar to interlaboratory comparison ζ-scores [46], which are the standard uncertainty normalized differences between measurements and an externally assigned value for the quantity. The problem with using actual ζ-scores is that they depend on having assigned values for the quantity independent of the measurements. Such values are not usually available for the quantities studied here, so any assigned value must be determined from the measurements themselves, and such 'consensus values' can be problematic [46]. The particular issue with h is whether the weighted mean $\bar{x}$ is the best assigned value for a quantity given all the available measurements. This is a reasonable assumption if the uncertainties are Normal, since then $\bar{x}$ from equation (3.4) is the maximum likelihood value for x [16]. If the uncertainties are not Normal, however, $\bar{x}$ may be far from maximum likelihood, so it is not clear if $\bar{x}$ is the best choice for the assigned value. Because of these issues, z was preferred over h in this study, but the h- and z-distributions are very similar. As can be seen from table 3, the fit quality and parameter values are comparable for h- and z-distributions, except the tails appear even heavier in h.

Table 4. Fitted Student's *t*-parameters for selected data, with number of quantities, measurements and comparison pairs.

	ν	σ	Quant.	Meas.	Pairs
Key	1.9 ± 0.1	1.12 ± 0.04	372	3714	20 308
Key Metrology	3.2 ± 0.2	0.94 ± 0.02	197	2030	12 070
Selected Metrology	9.9 ± 2.6	0.90 ± 0.03	156	575	948
Key Analytical	1.9 ± 0.2	1.62 ± 0.13	133	1238	5938
Selected Analytical	2.1 ± 0.3	1.39 ± 0.08	127	503	848
New Stable (since 2000)	7.6 ± 1.3	0.90 ± 0.03	357	1278	2478
BaBar/Belle stable	6.7 ± 2.1	0.91 ± 0.04	172	435	468
Other new stable	5.3 ± 0.8	0.79 ± 0.03	209	752	1395
Nuclear	2.0 ± 0.1	0.90 ± 0.02	1437	12 380	66 677
lifetimes	2.1 ± 0.2	1.30 ± 0.09	152	1560	9779
$u_x/x > 0.005$	2.2 ± 0.2	1.04 ± 0.06	125	759	3123
$u_x/x < 0.005$	2.8 ± 0.5	1.89 ± 0.22	110	772	3503
Constants	1.8 ± 0.2	0.89 ± 0.06	7	320	9098
Constants without G	3.2 ± 0.5	0.99 ± 0.11	6	231	5182

3.4. Selected data subsets

To further investigate the variance in the distributions for different types of measurements, several additional data subsets were examined and their parameters listed in table 4.

The Key Metrology data subset is for electrical, radioactivity, length, mass and other similar physical metrology standards. To see whether the most experienced national laboratories were more consistent, table 4 also lists Selected Metrology data from only the six national laboratories that reported the most Key Metrology measurements. These laboratories were PTB (Physikalisch-Technische Bundesanstalt, Germany), NMIJ (National Metrology Institute of Japan), NIST (National Institutes of Standards and Technology, USA), NPL (National Physical Laboratory, UK), NRC (National Research Council, Canada) and LNE (Laboratoire national de métrologie et d'essais, France). Similarly, Key Analytical chemistry data selected from the same national laboratories are also shown. These are for measurements such as the amount of mercury in salmon, PCBs in sediment or chromium in steel. The metrology measurements by the selected national laboratories do have much lighter tails with $\nu \sim 10$, but this is not the case for their analytical measurements where $\nu \sim 2$.

New Stable particle data have the lightest tail in table 2, but it is not clear if this is because the newer results have better determined uncertainties or are just more correlated. The trend in particle physics is for fewer but larger experiments, and more than a third of the newer Stable measurements were made by just two very similar experiments (Belle and BaBar), so the New Stable data are split into two groups in table 4. There is no significant difference between the Belle/BaBar and Other experiments data.

Nuclear lifetimes with small and large relative uncertainties were compared. They have similar tails, but the smaller uncertainty measurements appear to underestimate their uncertainty scales.

Measurements of Newton's gravitation constant are notoriously variable [14,47], so a dataset without G_N results was examined. The heavy tail is reduced, albeit with large uncertainty.

3.5. Relative uncertainty

The accuracy of uncertainty evaluations appears to be similar in all fields, but unsurprisingly there are notable differences in the relative sizes of the uncertainties. In particular, although individual physics measurements are not typically more reproducible than in medicine, they often have smaller relative uncertainty (i.e. uncertainty/value) as shown in figure 3.

Perhaps more importantly for discovery reproducibility, uncertainty improves more rapidly in physics than in medicine, as is shown in figure 4. This difference in rates of improvement reflects the difference between measurements that depend on steadily evolving technology versus those using stable

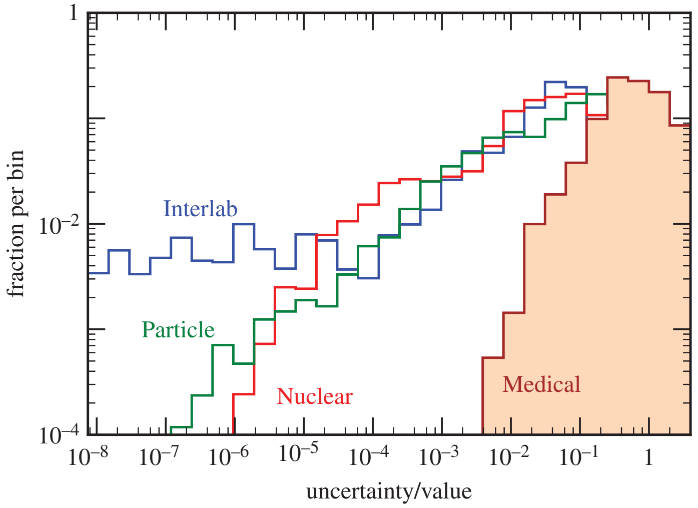

Figure 3. Distribution of the relative uncertainty for data from figure 1.

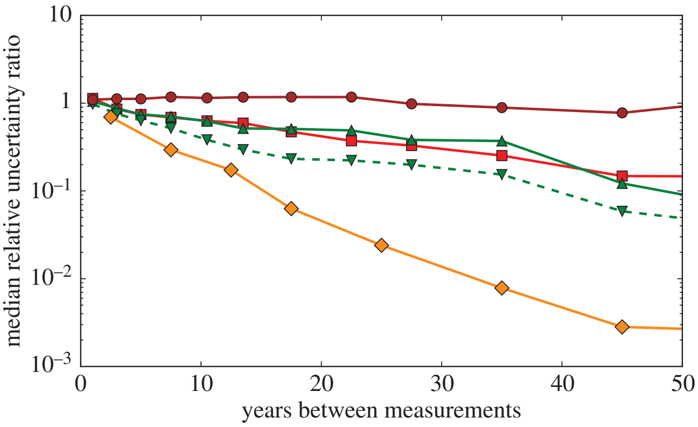

Figure 4. Median ratio of the relative uncertainties (newer/older) for measurements in each *z*-pair as a function of the years between the two measurements: Medical (brown circles), Particle (green point-up triangles), Nuclear (red squares), Stable (green dashed point-down triangles), Constants (orange diamonds).

methods that are limited by sample sizes and heterogeneity [48]. The expectation of reduced uncertainty in physics means that it is feasible to take a wait-and-see attitude towards new discoveries, since better measurements will quickly confirm or refute the new result. Measurement uncertainty in nuclear and particle physics typically improves by about a factor of 2 every 15 years. Constants data improve twice as fast, which is unsurprising since more effort is expected for more important quantities.

Physicists also tend not to make new measurements unless they are expected to be more accurate than previous measurements. In the datasets reported here, the median improvement in uncertainty of Nuclear measurements compared to the best previous measurement of the same quantity is $u_{\text{best}}/u_{\text{new}} = 2.0 \pm 0.3$, and the improvement factors for Constants, Particle and Stable measurements are 1.8 ± 0.3, 1.7 ± 0.2 and 1.3 ± 0.1, respectively. By contrast, Medical measurements typically have greater uncertainties than the best previous measurements, with median $u_{\text{best}}/u_{\text{new}} = 0.62 \pm 0.03$. This is an understandable consequence of different uncertainty to cost relationships in physics and medicine. Study population size is a major cost driver in medical research, so reducing the uncertainty by a factor of two can cost almost four times as much, which is rarely the case in physics.

3.6. Expectations and correlations

Prior expectations exist for most measurements reported here except for the Interlab data. Such expectations may suppress heavy tails by discouraging publication of the anomalous results that

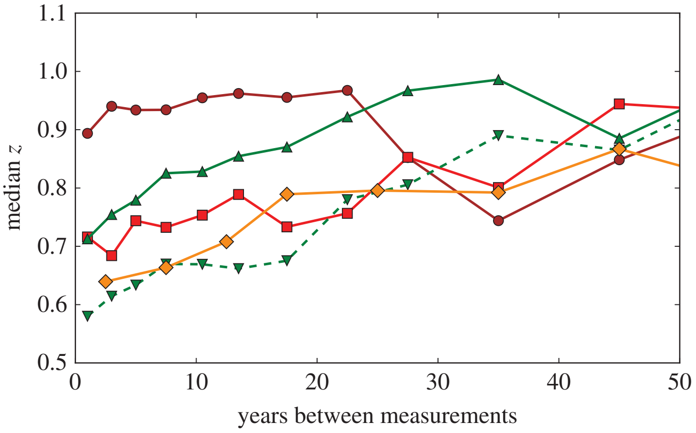

Figure 5. Median z-value as a function of time difference between the two measurements in each z-pair: Medical (brown circles), Particle (green point-up triangles), Nuclear (red squares), Constants (orange diamonds) and Stable (green dashed point-down triangles).

populate the tails, since before publishing a result dramatically different from prior results or theoretical expectations, researchers are likely to make great efforts to ensure that they have not made a mistake. Journal editors, referees and other readers also ask tough questions of such results, either preventing publication or inducing further investigation. For example, initial claims [49,50] of 6σ evidence for faster-than-light neutrinos and cosmic inflation did not survive to actual publication [51,52].

Figure 5 shows that physics (Particle, Nuclear and Constants) measurements are more likely to agree if the difference in their publication dates is small. Such 'bandwagon effects' [20,36] are not observed in the Medical data, and they are irrelevant for Interlab quantities which are usually measured almost simultaneously. These correlations imply that measurements are biased either by expectations or common methodologies. Such correlations might explain the small (less than 1) values of σ_x for Nuclear, Particle and Constants data, or it could be that researchers in these fields simply tend to overestimate the scale of their uncertainties [53]. Removing expectation biases from the physics data would probably make their tails heavier.

Although Interlab data are not supposed to have any expectation biases, they are subject to methodological correlations due to common measurement models, procedures and types of instrumentation, so even their tails would probably increase if all measurements could be made truly independent.

4. Discussion

4.1. Comparison with earlier studies

In a famous dispute with Cauchy in 1853, eminent statistician Irénée-Jules Bienaymé ridiculed the idea that any sensible instrument had Cauchy uncertainties [54]. A century later, however, Harold Jeffreys noted that systematic errors may have a significant Cauchy component, and that the scale of the uncertainty contributed by systematic effects depends on the size of the random errors [55].

The results of this study agree with earlier research that also observed Student's t-tails, but only looked at a handful of subatomic or astrophysics quantities up to $z \sim 5-10$ [16,19,56–58]. Unsurprisingly, the tails reported here are mostly heavier than those reported for repeated measurements made with the same instrument ($\nu \sim 3-9$) [59–61], which should be closer to Normal as they are not independent and share most systematic effects.

Instead of Student's t tails, exponential tails have been reported for several nuclear and particle physics datasets [15,17,18,20], but in all cases some measurements were excluded. For example, the largest of these studies [20] looked at particle data (315 quantities and 53 322 pairs) using essentially the same method as this paper, but rejected the 20% of the data that gave the largest contributions to the χ^2 for each quantity, suppressing the heaviest tails. Despite this data selection, all these studies have supra-exponential heavy tails for $z \gtrsim 5$, and so are qualitatively consistent with the results of this paper. It is possible that averaging different quantities with exponential tails might produce apparent power laws [62], but this would require wild variations in the accuracy of the uncertainty estimates.

Instead of looking directly at the shapes of the measurement consistency distributions, Hedges [10] compared particle physics and psychology results and found them to have similar compatibility, with typically almost half of the quantities in both fields having statistically significant disagreements.

Thompson & Ellison [63] reported substantial amounts of 'dark uncertainty' in chemical analysis interlaboratory comparisons. Uncertainty is 'dark' if it does not appear as part of the known contributions to the uncertainty of individual measurements, but is inferred to exist because the dispersion of measured values is greater than expected based on the reported uncertainties. For example, six (21%) of 28 BIPM Key Comparisons studied had ratios ($\bar{s}_{\text{exp}}/s_{\text{obs}}$) of expected to observed standard deviations less than 0.2. This agrees with the key analytical results in table 4 (which include some of the same key comparisons). For sample sizes matching the 28 comparisons, 20% of samples drawn from a $\nu = 2, \sigma = 1.4$ Student's t-distribution would be expected to have $\bar{s}_{\text{exp}}/s_{\text{obs}} < 0.2$. Pavese has also emphasized the large number of discrepant results in Key comparisons [64].

The Open Science Collaboration (OSC) recently replicated 100 studies in psychology [65], providing some of the most direct evidence yet for poor scientific reproducibility. Using the OSC study's supplementary information, z can be calculated for 87 of the reported original/replication measurement pairs, and 27 (31%) disagree by more than 2σ, and 2 (2.3%) by more than 5σ. This rate of disagreements is inconsistent with selection bias acting on a Normal distribution unless the more than 5σ data are excluded, but can be explained by selection-biased Student's t-data with $\nu \sim 3$, consistent with the medical data reported in table 2.

4.2. How measurements fail

When a measurement turns out to be wrong, the reasons for this failure are often unknown, or at least unpublished, so it is interesting to look at examples where the causes were later understood or can be inferred.

For medical research, heterogeneity in methods or populations is a major source of variance. The largest inconsistency in the Medical dataset is in a comparison of fever rates after acellular versus whole-cell pertussis vaccines [66]. The large variance can probably be explained by significant differences among the study populations and especially in how minor adverse events were defined and reported.

The biggest z-values in the particle data come from complicated multi-channel partial wave analyses of strong scattering processes, where many dozens of quantities (particle masses, widths, helicities, etc.) are simultaneously determined. Significant correlations often exist between the fitted values of the parameters but are not always clearly reported, and evaluations may not always include the often large uncertainties from choices in data and parametrization.

The largest disagreement in the Interlab data appears to be an obvious mistake. In a comparison of radioactivity in water [67], one laboratory reported an activity of $139\,352 \pm 0.82\,\text{Bq}\,\text{kg}^{-1}$ when the true value was about 31. Even without knowing the expected activity, the unreasonably small fractional uncertainty should probably have flagged this result. Such gross errors can produce almost-Cauchy deviations. For example, if the numerical result of a measurement is simply considered as an infinite bit string, then any 'typographical' glitch that randomly flips any bit with equal probability will produce deviations with a $1/x$ distribution.

One can hope that the best research will not be sloppy, but not even the most careful scientists can avoid all unpleasant surprises. In 1996, a team from PTB (the National Metrological Institute of Germany) reported a measurement of G_N that differed by 50σ from the accepted value; it took 8 years to track down the cause—a plausible but erroneous assumption about their electrostatic torque transmitter unit [68]. A 6.5σ difference between the CODATA2006 and CODATA2010 fine-structure constant values was due to a mistake in the calculation of some eighth-order terms in the theoretical value of the electron anomalous magnetic moment [69]. A 1999 determination [70] of Avogadro's number by a team from Japan's National Research Laboratory of Metrology using the newer X-ray crystal density method was off by approximately 9σ due to subtle silicon inhomogeneities [71]. In an interlaboratory comparison measuring PCB contamination in sediments, the initial measurement by BAM (the German Federal Institute for Materials Research and Testing) disagreed by many standard uncertainties, but this was later traced to cross-contamination in sample preparation [72]. Several nuclear half-lives measured by the US National Institute for Standards and Technology were known for some years to be inconsistent with other measurements; it was finally discovered that a NIST sample positioning ring had been slowly slipping over 35 years of use [73].

Often discrepancies are never understood and are simply replaced by newer results. For example, despite bringing in a whole new research team to go over every component and system, the reason for a discordant NIST measurement of Planck's constant was never found, but newer measurements by the same group were not anomalous [74].

4.3. Causes of heavy tails

Heavy tails have many potential causes, including bias [7], overconfident uncertainty underestimates [75] and uncertainty in the uncertainties [17], but it is not immediately obvious how these would produce the observed t-distributions with so few degrees of freedom.

Even when the uncertainty u is evaluated from the standard deviation of multiple measurements from a Normal distribution so that Student's t-distribution would be expected, there are typically so many measurements that ν should be much larger than what is observed. Exceptions to this are when calibration uncertainties dominate, since often only a few independent calibration points are available, or when uncertainties from systematic effects are evaluated by making a few variations to the measurements, but these cannot explain most of the data.

Any reasonable publication bias applied to measurements with Gaussian uncertainties cannot create very heavy tails, just a distorted distribution with Gaussian tails—to produce one false published 5σ result would require bias strong enough to reject millions of studies. Underestimating σ does not produce a heavy tail, only a broader Normal z-distribution. Mixing multiple Normal distributions does not naturally produce almost-Cauchy distributions, except in special cases such as the ratio of two zero-mean Gaussians.

The heavy tails are not caused by poor older results. The heaviest-tailed data in figure 1 are actually the newest—93% of the interlaboratory data are less than 16 years old—and eliminating older results taken prior to the year 2000 does not reduce the tails for most data as shown in table 2.

Intentionally making up results, i.e. fraud, could certainly produce outliers, but this is unlikely to be a significant problem here. Since most of the data were extracted from secondary meta-analyses (e.g. Review of Particle Properties, Table of Radionuclides and Cochrane systematic reviews), results withdrawn for misconduct prior to the time of the review would probably be excluded. One meta-analysis in the medical dataset does include studies that were later shown to be fraudulent [76], but the fraudulent results actually contribute slightly less than average to the overall variance among the results for that meta-analysis.

4.4. Modelling

Modelling the heavy tails may help us understand the observed distributions. One way is to assume that the measurement values are normally distributed with standard deviation t that is unknown but which has a probability distribution $f(t)$ [15,17–19,77]. The measured value x is then expected to have a probability distribution

$$\mathcal{P}(x) = \int_0^\infty \mathrm{d}t f(t) \frac{1}{\sqrt{2\pi}\,t} \mathrm{e}^{-x^2/(2t^2)}. \tag{4.1}$$

This is essentially a Bayesian estimate with prior $f(t)$ and a Normal likelihood with unknown variance. If the uncertainties are accurately evaluated and Normal with variance σ^2, $f(t)$ will be a narrow peak at $t = \sigma$. Assuming that $f(t)$ is a broad Normal distribution leads to exponential tails [17] for large z.

To generate Student's t-distributions, $f(t)$ must be a scaled inverse χ^2- (or Gamma) distribution in t^2 [19,77]. This works mathematically, but why would variations in σ for independent measurements have such a distribution?

Heavy tails can only be generated by effects that can produce a wide range of variance, so we must model how consistency testing is used by researchers to constrain such effects. Consistency is typically tested using a metric such as the calculated χ^2-statistic for the agreement of N measurements x_i [43]

$$\chi_c^2(x, u) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{u_i^2}, \tag{4.2}$$

where $\bar{x}$ is the x_i weighted mean and u_i are the standard uncertainties reported by the researchers. For accurate standard uncertainties, χ_c^2 will have a χ^2- probability distribution with $\nu = N - 1$. If, however, the reported uncertainties are incorrect and the true standard uncertainties are tu_i, then it will be $\chi^2_{\text{true}}(x, tu) = \chi_c^2(x, u)/t^2$ that is χ^2-distributed.

Researchers will probably search for problems if different consistency measurements have a poor $\chi_c^2(x, u)$, which typically means $\chi_c^2(x, u) > \nu$. The larger an unknown systematic error is, the more likely it is to be detected and either corrected or included in the reported uncertainty, so published results typically have $\chi_c^2(x,u) \sim \nu$. Since $\chi_c^2(x,u)/t^2$ is expected to have a χ^2-distribution, a natural prior for t^2 is indeed the scaled inverse χ^2-distribution needed to generate Student's t-distributions from equation (4.1).

More mechanistically, it could be assumed that a normally distributed systematic error will be missed by N_m independent measurements if their $\chi^2(u) = (t^2/u^2)\chi^2(t)$ is less than some threshold $\chi^2_{\text{max}} \sim \nu = N_m - 1$. If the distribution of all possible systematic effects is $P_0(t)$, then the probability distribution for the unfound errors will be

$$f(t;\nu) = P_0(t) F\left(\frac{\chi^2_{\text{max}};\nu}{t^2}\right), \tag{4.3}$$

where F is the cumulative χ^2-distribution. $P_0(t)$ is unknown, but a common Bayesian scale-invariant choice is $P_0(t) \propto 1/t^\alpha$, with $\alpha > 0$.

Using this model with the reported uncertainty σ as the lower integration bound, the curve generated from equations (4.1) and (4.3) is very close to a $\nu = N_m - 1 + \alpha$ Student's t-distribution. The observed small values for ν mean that both N_m and α must be small. Making truly independent consistency tests is difficult, so it is not surprising that the effective number of checks (N_m) is usually small.

This model is plausible, but why are systematic effects consistent with a $P_0(t) \propto 1/t^\alpha$ power-law size distribution?

4.5. Complex systems

Scientific measurements are made by complex systems of people and procedures, hardware and software, so one would expect the distribution of scientific errors to be similar to those produced by other comparable systems.

Power-law behaviour is ubiquitous in complex systems [78], with the cumulative distributions of observed sizes (s) for many effects falling as $1/s^\alpha$, and these heavy tails exist even when the system has been designed and refined for optimal results.

A Student's t-distribution has cumulative tail exponent $\alpha = \nu$, and the values for ν reported here are consistent with power-law tails observed in other designed complex systems. The frequency of software errors typically has a cumulative power-law tail corresponding to small $\nu \sim 2-3$ [79], and in scientific computing these errors can lead to quantitative discrepancies orders of magnitude greater than expected [80]. The size distribution of electrical power grid failures has $\nu \sim 1.5-2$ [81], and the frequency of spacecraft failures has $\nu \sim 0.6-1.3$ [82]. Even when designers and operators really, really want to avoid mistakes, they still occur: the severity of nuclear accidents falls off only as $\nu \sim 0.7$ [83], similar to the power-laws observed for the sizes of industrial accidents [84] and oil spills [85]. Some complex medical interventions have power-law distributed outcomes with $\nu \sim 3-4$ [86].

Combining the observed power-law responses of complex systems with the power-law constraints of consistency checking for systematic effects discussed in §4.4, leads naturally to the observed consistency distributions with heavy power-law tails. There are also several theoretical arguments that such distributions should be expected.

A systematic error or mistake is an example of a risk analysis incident, and power-law distributions are the maximal entropy solutions for such incidents when there are multiple nonlinear interdependent causes [85], which is often the case when things go wrong in research.

Scientists want to make the best measurements possible with the limited resources they have available, so scientific research endeavours are good examples of highly structured complex systems designed to optimize outcomes in the presence of constraints. Such systems are expected to exhibit 'highly optimized tolerance' [87,88], being very robust against designed-for uncertainties, but also hypersensitive to unanticipated effects, resulting in power-law distributed responses. Simple continuous models for highly optimized tolerant systems are consistent with the heavy tails observed in this study. These models predict that $\alpha \sim 1 + 1/d$ [88,89], where $d(> 0)$ is the effective dimensionality of the system, but larger values of α arise when some of the resources are used to avoid large deviations [89], e.g. spending time doing consistency checks.

4.6. How can heavy tails be reduced?

If one believes that mistakes can be eliminated and all systematic errors found if we just work hard enough and apply the most rigorous methodological and statistical techniques, then results from the best scientists should not have heavy tails. Such a belief, however, is not consistent with the experienced challenges of experimental science, which are usually hidden in most papers reporting scientific measurements [4,90]. As Beveridge famously noted [91], often everyone else believes an experiment more than the experimenters themselves. Researchers always fear that there are unknown problems with their work, and traditional error analysis cannot 'include what was not thought of' [47].

It is not easy to make accurate *a priori* identifications of those measurements that are so well done that they avoid having almost-Cauchy tails. Expert judgement is subject to well-known biases [92], and obvious criteria to identify better measurements may not work. For example, the Open Science Collaboration found that researchers' experience or expertise did not significantly correlate with the reproducibility of their results [65]—the best predictive factor was simply the statistical significance of the original result. The best researchers may be better at identifying problems and not making mistakes, but they also tend to choose the most difficult challenges that provide the most opportunities to go wrong.

Reducing heavy tails is challenging because complex systems exhibit scale invariant behaviour such that reducing the size of failures does not significantly change the shape of their distribution. Improving sensitivity makes previously unknown small systematic issues visible so they can be corrected or included in the total uncertainty. This improvement reduces σ, but even smaller systematic effects now become significant and tails may even become heavier and ν smaller. Comparing figures 1 and 3, it appears that data with higher average relative uncertainty tend to have heavier tails. This relationship between relative uncertainty and measurement dispersion is reminiscent of the empirical Horwitz power law in analytical chemistry [93], where the relative spread of interlaboratory measurements increases as the required sensitivity gets smaller, and of Taylor's Law in ecology, where the variance grows with sample size so that the uncertainty on the mean does not shrink as $1/\sqrt{N}$ [94].

In principle, statistical errors can be made arbitrarily small by taking enough data, and ν can be made arbitrarily large by making enough independent consistency checks, but researchers have only finite time and resources so choices must be made. Taking more consistency check data limits the statistical uncertainty, since it is risky to treat data taken under different conditions as a single dataset. Consistency checks are never completely independent since it is impossible for different measurements of the same quantity not to share any people, methods, apparatus, theory or biases, so researchers must decide what tests are reasonable. The observed similar small values for ν may reflect similar spontaneous and often unconscious cost–benefit analyses made by researchers.

The data showing the lightest tail reported here (in table 4) may provide some guidance and caution. The high quality of the selected metrology standards measurements by leading national laboratories shows that heavy tails can be reduced by collaboratively taking great care to ensure consistency by sharing methodology and making regular comparisons. There are, however, limits to what can be achieved, as illustrated by the much heavier tail of the analytical standards measured by the same leading labs. Secondly, consistency is easier than accuracy. Interlaboratory comparisons typically take place over relatively short periods of time, with participating institutions using the best standard methods available at that time. Biases in the standard methods may only be later discovered when new methods are introduced. For example, work towards a redefinition of the kilogram and the associated development of new silicon atom counting technology revealed inconsistencies with earlier watt-balance measurements, and this has driven improvements in both methods [74]. Finally, selection bias that hides anomalous results is hard to eliminate. For one metrology key comparison, results from one quantity were not published because some laboratories reported 'incorrect results' [95].

Reducing tails is particularly challenging for measurements where the primary goal is improved sensitivity that may lead to new scientific understanding. By definition, a better measurement is not an identical measurement, and every difference provides room for new systematic errors, and every improvement that reduces the uncertainty makes smaller systematic effects more significant. Frontier measurements are always likely to have heavier tails.

5. Conclusion

Published scientific measurements typically have non-Gaussian almost-Cauchy $\nu \sim 2-4$ Student's *t*-error distributions, with up to 10% of results in disagreement by greater than 5σ. These heavy tails occur in even the most careful modern research, and do not appear to be caused by selection bias, old inaccurate data, or sloppy measurements of uninteresting quantities. For even the best scientists working on well-understood measurements using similar methodology, it appears difficult to achieve consistency better than $\nu \sim 10$, with about 0.1% of results expected to be greater than 5σ outliers, a rate a thousand times higher than for a Normal distribution. These may, however, be underestimates. Because of selection/confirmation biases and methodological correlations, historical consistency can only set lower bounds on heavy tails—multiple measurements may all agree but all be (somewhat) wrong.

The effects of unknown systematic problems are not completely unpredictable. Scientific measurement is a complex process and the observed distributions are consistent with unknown systematics following the low-exponent power-laws that are theoretically expected and experimentally observed for fluctuations and failures in almost all complex systems.

Researchers do determine the scale of their uncertainties with fair accuracy, with the scale of medical uncertainties ($\sigma_x \sim 1$) slightly more consistent with the expected value ($\sigma_x = 1$) than in physics ($\sigma_x \sim 0.7-0.8$). Medical and physics research have comparable reproducibility in terms of how well different studies agree within their uncertainties, consistent with a previous comparison of particle physics with social sciences [10]. Medical research may have slightly lighter tails, while physics results typically have better relative uncertainty and greater statistical significance.

Understanding that error distributions are often almost-Cauchy should encourage use of t-based [96], median [97] and other robust statistical methods [98], and supports choosing Student's t [99] or Cauchy [100] priors in Bayesian analysis. Outlier-tolerant methods are already common in modern meta-analysis, so there should be little effect on accepted values of quantities with multiple published measurements, but this better understanding of the uncertainty may help improve methods and encourage consistency.

False discoveries are more likely if researchers apply Normal conventions to almost-Cauchy data. Although much abused, the historically common use of $p < 0.05$ as a discovery criterion suggests that many scientists would like to be wrong less than 5% of the time. If so, the results reported here support the nominal 5σ discovery rule in particle physics, and may help discussion of more rigorous significance criteria in other fields [101–103].

This study should help researchers better understand the uncertainties in their measurements, and may help decision-makers and the public better interpret the implications of scientific research [104]. If nothing else, it should also remind everyone to never use Normal/Gaussian statistics when discussing the likelihood of extreme results.

Data accessibility. The sources for all data analysed are listed at http://dx.doi.org/10.5061/dryad.jb3mj [37].
Competing interests. I have no competing interests.
Funding. Financial support came from the University of Toronto.
Acknowledgements. I thank the students of the University of Toronto Advanced Undergraduate Physics Lab for inspiring consideration of realistic experimental expectations, and the University of Auckland Physics Department for their hospitality during very early stages of this work. I am grateful to D. Pitman for patient and extensive feedback, to R. Bailey for his constructive criticism, to R. Cousins, D. Harrison, J. Rosenthal and P. Sinervo for useful suggestions and discussion, and to M. Cox for many helpful comments on the manuscript.

References

1. McNutt M. 2014 Journals unite for reproducibility. *Science* **346**, 679. (doi:10.1126/science.aaa1724)
2. Conrad J. 2015 Reproducibility: don't cry wolf. *Nature* **523**, 27–28. (doi:10.1038/523027a)
3. Rosenfeld AH. 1968 Are there any far-out mesons or baryons? In *Meson spectroscopy* (eds C Baltay, AH Rosenfeld), pp. 455–483. New York, NY: W. A. Benjamin.
4. Franklin A. 2013 *Shifting standards: experiments in particle physics in the twentieth century*. Pittsburgh, PA: University of Pittsburgh Press.
5. Cousins RD. 2014 The Jeffreys-Lindley paradox and discovery criteria in high energy physics. *Synthese* 1–38. (doi:10.1007/s11229-014-0525-z)
6. Dorigo T. 2015 Extraordinary claims: the 0.000029% solution. *EPJ Web Conf.* **95**, 02003. (doi:10.1051/epjconf/20149502003)
7. Ioannidis JPA. 2005 Why most published research findings are false. *PLoS Med.* **2**, 696–701. (doi:10.1371/journal.pmed.0020124)
8. Nakagawa S, Cuthill IC. 2007 Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* **82**, 591–605. (doi:10.1111/j.1469-185X.2007.00027.x)
9. Cumming G. 2014 The new statistics: why and how. *Psychol. Sci.* **25**, 7–29. (doi:10.1177/0956797613504966)
10. Hedges LV. 1987 How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *Am. Psychol.* **42**, 443–455. (doi:10.1037/0003-066X.42.5.443)
11. Porter FC. 2006 The significance of HEP observations. *Int. J. Mod. Phys. A* **21**, 5574–5582. (doi:10.1142/S0217751X06034768)
12. Youden WJ. 1972 Enduring values. *Technometrics* **14**, 1–11. (doi:10.2307/1266913)
13. Stigler SM. 1977 Do robust estimators work with real data. *Ann. Stat.* **5**, 1055–1098. (doi:10.1214/aos/1176343997)
14. Speake C, Quinn T. 2014 The search for Newton's constant. *Phys. Today* **67**, 27–33. (doi:10.1063/PT.3.2447)
15. Bukhvostov AP. 1973 *On the statistical meaning of experimental data*. Report no. 45. Leningrad, Russia: Leningrad Nuclear Physics Institute.
16. Roos M, Hietanen M, Luoma J. 1975 A new procedure for averaging particle properties. *Physica Fennica* **10**, 21–33.
17. Shlyakhter AI. 1994 An improved framework for uncertainty analysis: accounting for unsuspected errors. *Risk Anal.* **14**, 441–447. (doi:10.1111/j.1539-6924.1994.tb00262.x)
18. Bukhvostov AP. 1997 On the probability distribution of the experimental results. (http://arxiv.org/abs/hep-ph/9705387)
19. Hanson KM. 2007 Lessons about likelihood functions from nuclear physics. *AIP Conf. Proc.* **954**, 458–467. (doi:10.1063/1.2821298)
20. Jeng M. 2007 Bandwagon effects and error bars in particle physics. *Nucl. Instrum. Meth. A* **571**, 704–708. (doi:10.1016/j.nima.2006.11.024)
21. Joint Committee for Guides in Metrology. 2008 Guide to the expression of uncertainty in measurement. JCGM 100.
22. Sinervo PK. 2003 Definition and treatment of systematic uncertainties in high energy physics and astrophysics. In *Proc. Conf. on Statistical Problems in Particle Physics, Astrophysics and*

Cosmology (*PHYSTAT 2003*), *Stanford, CA, 8–11 September*, pp. 122–129.

23. Carlson CE. 2015 The proton radius puzzle. *Prog. Part. Nucl. Phys.* **82**, 59–77. (doi:10.1016/j.ppnp.2015.01.002)
24. Shields D. 2015 Giving credit where credit is due. *Geotechn. Instrum. News* **December**, 33–34.
25. Bravin E *et al.* 1998 The influence of train leakage currents on the LEP dipole field. *Nucl. Instrum. Meth. A* **417**, 9–15. (doi:10.1016/S0168-9002(98)00020-5)
26. Colclough AR. 1987 Two theories of experimental error. *J. Res. Nat. Bur. Stand.* **92**, 167–185. (doi:10.6028/jres.092.016)
27. Barlow R. 2002 Systematic errors: facts and fictions. (http://arxiv.org/abs/hep-ex/0207026)
28. Pavese F. 2009 About the treatment of systematic effects in metrology. *Measurement* **42**, 1459–1462. (doi:10.1016/j.measurement.2009.07.017)
29. Attivissimo F, Cataldo A, Fabbiano L, Giaquinto N. 2011 Systematic errors and measurement uncertainty: an experimental approach. *Measurement* **44**, 1781–1789. (doi:10.1016/j.measurement.2011.07.011)
30. Dorsey NE. 1944 The velocity of light. *Trans. Am. Philos. Soc.* **34**, 1–110. (doi:10.2307/1005532)
31. Pommé S. 2016 When the model doesn't cover reality: examples from radionuclide metrology. *Metrologia* **53**, S55–S64. (doi:10.1088/0026-1394/53/2/S55)
32. Cochrane Collaboration. 2013 Cochrane database of systematic reviews. See http://www.thecochranelibrary.com.
33. Beringer J *et al.* 2012 Review of Particle Physics. *Phys. Rev. D* **86**, 010001. (doi:10.1103/PhysRevD.86.010001)
34. Beringer J *et al.* 2013 Review of Particle Physics 2013 partial update for the 2014 edition. See http://pdg.lbl.gov/2013/tables/contents_tables.html.
35. Bé MM *et al.* 2013 Table of Radionuclides (comments on evaluation). Bureau international des poids et mesures BIPM-5. See http://bit.ly/2fuoaU3.
36. Baker RD, Jackson D. 2013 Meta-analysis inside and outside particle physics: two traditions that should converge? *Res. Synth. Meth.* **4**, 109–124. (doi:10.1002/jrsm.1065)
37. Bailey DC. 2017 Data from: Not Normal: the uncertainties of scientific measurements. *Dryad Digital Repository*. (http://dx.doi.org/10.5061/dryad.jb3mj).
38. Beringer J *et al*. 2013 archives and errata for the Review of Particle Physics. See http://pdg.lbl.gov/2013/html/rpp archives.html.
39. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. 2003 Measuring inconsistency in meta-analyses. *Brit. Med. J.* **327**, 557–560. (doi:10.1136/bmj.327.7414.557)
40. Martín Andrés A, Álvarez Hernández M. 2014 Two-tailed approximate confidence intervals for the ratio of proportions. *Stat. Comput.* **24**, 65–75. (doi:10.1007/s11222-012-9353-5)
41. Joint Committee for Guides in Metrology. 2012 International vocabulary of metrology—basic and general concepts and associated terms (VIM). JCGM 200. See http://www.bipm.org/en/publications/guides/vim.html.
42. Barlow R. 2003 Asymmetric errors. In *Proc. Conf. on Statistical Problems in Particle Physics, Astrophysics and Cosmology* (*PHYSTAT 2003*), *Stanford, CA, 8–11 September*, pp. 250–255.
43. Bohm G, Zech G. 2010 *Introduction to statistics and data analysis for physicists*. Verlag Deutsches Elektronen-Synchrotron. (doi:10.3204/DESY-BOOK/statistics)
44. Bureau International des Poids et Mesures. 2014 The BIPM key comparison database. See http://kcdb.bipm.org.
45. Willink R. 2004 Approximating the difference of two t-variables for all degrees of freedom using truncated variables. *Aust. N. Z. J. Stat.* **46**, 495–504. (doi:10.1111/j.1467-842X.2004.00346.x)
46. International Standards Organization. 2005 *ISO 13528:2005 Statistical methods for use in proficiency testing by interlaboratory comparisons*. Geneva, Switzerland: ISO.
47. Faller JE. 2014 Precision measurement, scientific personalities and error budgets: the *sine quibus non* for big *G* determinations. *Phil. Trans. R. Soc. A* **372**, 20140023. (doi:10.1098/rsta.2014.0023)
48. Peterson D. 2015 All that is solid: bench-building at the frontiers of two experimental sciences. *Am. Sociol. Rev.* **80**, 1201–1225. (doi:10.1177/0003122415607230)
49. Adam T *et al.* 2011 Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. (http://arxiv.org/abs/1109.4897v1)
50. Ade PAR *et al.* 2014 Detection of B-mode polarization at degree angular scales by BICEP2. (http://arxiv.org/abs/1403.3985v1)
51. Adam T *et al.* 2012 Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. *J. High Energy Phys.* **1210**, 093. (doi:10.1007/JHEP10(2012)093)
52. Ade PAR *et al.* 2014 Detection of B-mode polarization at degree angular scales by BICEP2. *Phys. Rev. Lett.* **112**, 241101. (doi:10.1103/PhysRevLett.112.241101)
53. Nachman B, Rudelius T. 2012 Evidence for conservatism in LHC SUSY searches. *Eur. Phys. J. Plus* **127**, 157. (doi:10.1140/epjp/i2012-12157-0)
54. Bienaymé M. 1853 Considérations a l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. *C. R. Acad. Sci.* **37**, 309–324.
55. Jeffreys H. 1961 *The theory of probability*, 3rd edn. Oxford, UK: Oxford University Press.
56. Chen G, Gott JR, Ratra B. 2003 Non-Gaussian error distribution of Hubble constant measurements. *Publ. Astron. Soc. Pac.* **115**, 1269–1279. (doi:10.1086/379219)
57. Crandall S, Houston S, Ratra B. 2015 Non-Gaussian error distribution of ^{7}Li abundance measurements. *Mod. Phys. Lett. A* **30**, 1550123. (doi:10.1142/S0217732315501230)
58. Crandall S, Ratra B. 2015 Non-Gaussian error distributions of LMC distance moduli measurements. *Astrophys. J.* **815**, 87. (doi:10.1088/0004-637X/815/2/87)
59. Jeffreys H. 1938 The law of error and the combination of observations. *Phil. Trans. R. Soc. Lond. A* **237**, 231–271. (doi:10.1098/rsta.1938.0008)
60. Jeffreys H. 1939 The law of error in the Greenwich variation of latitude observations. *Mon. Not. R. Astron. Soc.* **99**, 703–709. (doi:10.1093/mnras/99.9.703)
61. Dzhun' IV. 2012 Distribution of errors in multiple large-volume observations. *Meas. Tech.* **55**, 393–396. (doi:10.1007/s11018-012-9970-6)
62. Anderson R. 2001 The power law as an emergent property. *Mem. Cogn.* **29**, 1061–1068. (doi:10.3758/BF03195767)
63. Thompson M, Ellison SLR. 2011 Dark uncertainty. *Accredit. Qual. Assur.* **16**, 483–487. (doi:10.1007/s00769-011-0803-0)
64. Pavese F. 2015 Key comparisons: the chance for discrepant results and some consequences. *Acta IMEKO* **4**, 38–47. (doi:10.21014/acta imeko.v4i4.267)
65. Open Science Collaboration. 2015 Estimating the reproducibility of psychological science. *Science* **349**, 6251. (doi:10.1126/science.aac4716)
66. Zhang L, Prietsch SO, Axelsson I, Halperin SA. 2012 Acellular vaccines for preventing whooping cough in children. *Cochrane Database Syst. Rev.* CD001478. (doi:10.1002/14651858.CD001478.pub5)
67. Terrestrial Environment Laboratory, IAEA Seibersdorf. 2010 *ALMERA proficiency test determination of naturally occurring radionuclides in phosphogypsum and water* (*IAEA-CU-2008-04*). Austria: International Atomic Energy Agency IAEA/AQ/15. See http://www-pub.iaea.org/MTCD/publications/PDF/IAEA-AQ-15_web.pdf.
68. Michaelis W, Melcher J, Haars H. 2004 Supplementary investigations to PTB's evaluation of G. *Metrologia* **41**, L29–L32. (doi:10.1088/0026-1394/41/6/L01)
69. Mohr PJ, Taylor BN, Newell DB. 2012 CODATA recommended values of the fundamental physical constants: 2010. *J. Phys. Chem. Ref. Data* **41**, 043109. (doi:10.1063/1.4724320)
70. Fujii K, Tanaka M, Nezu Y, Nakayama K, Fujimoto H, De Bievre P, Valkiers S. 1999 Determination of the Avogadro constant by accurate measurement of the molar volume of a silicon crystal. *Metrologia* **36**, 455–464. (doi:10.1088/0026-1394/36/5/7)
71. De Bievre P *et al.* 2001 A reassessment of the molar volume of silicon and of the Avogadro constant. *IEEE Trans. Instrum. Meas.* **50**, 593–597. (doi:10.1109/19.918199)
72. Schantz M, Wise S. 2004 CCQM-K25: determination of PCB congeners in sediment. *Metrologia* **41**, 08001. (doi:10.1088/0026-1394/41/1A/08001)
73. Pommé S. 2015 The uncertainty of the half-life. *Metrologia* **52**, S51–S65. (doi:10.1088/0026-1394/52/3/S51)
74. Gibney E. 2015 Experiments to redefine kilogram converge at last. *Nature* **526**, 305–306. (doi:10.1038/526305a)
75. Henrion M, Fischhoff B. 1986 Assessing uncertainty in physical constants. *Am. J. Phys.* **54**, 791–798. (doi:10.1119/1.14447)
76. Carlisle J, Pace N, Cracknell J, Møller A, Pedersen T, Zacharias M. 2013 (5) What should the Cochrane Collaboration do about research that is, or might be, fraudulent? *Cochrane Database Syst. Rev*. (doi:10.1002/14651858.ED000060)
77. Dose V, Von Der Linden W. 1999 Outlier tolerant parameter estimation. In *Maximum Entropy and Bayesian Methods: Garching, Germany 1998*. Fundamental Theories of Physics,

vol. 105, pp. 47–56. The Netherlands: Springer.

78. Clauset A, Shalizi CR, Newman MEJ. 2009 Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703. (doi:10.1137/070710111)

79. Hatton L. 2012 Defects, scientific computation and the scientific method. *IFIP Adv. Inform. Commun. Technol.* **377**, 123–138. (doi:10.1007/978-3-642-32677-6_8)

80. Hatton L, Roberts A. 1994 How accurate is scientific software. *IEEE Trans. Softw. Eng.* **20**, 785–797. (doi:10.1109/32.328993)

81. Dobson I, Carreras BA, Lynch VE, Newman DE. 2007 Complex systems analysis of series of blackouts: cascading failure, critical points, and self-organization. *Chaos* **17**, 026103. (doi:10.1063/1.2737822)

82. Karimova LM, Kruglun OA, Makarenko NG, Romanova NV. 2011 Power law distribution in statistics of failures in operation of spacecraft onboard equipment. *Cosm. Res.* **49**, 458–463. (doi:10.1134/S0010952511040058)

83. Sornette D, Maillart T, Kroeger W. 2013 Exploring the limits of safety analysis in complex technological systems. *Int. J. Disaster Risk Reduct.* **6**, 59–66. (doi:10.1016/j.ijdrr.2013.04.002)

84. Lopes AM, Tenreiro Machado JA. 2015 Power law behavior and self-similarity in modern industrial accidents. *Int. J. Bifurc. Chaos* **25**, 1550004. (doi:10.1142/S0218127415500042)

85. Englehardt J. 2002 Scale invariance of incident size distributions in response to sizes of their causes. *Risk Anal.* **22**, 369–381. (doi:10.1111/0272-4332.00016)

86. Burton C. 2012 Heavy tailed distributions of effect sizes in systematic reviews of complex interventions. *PLoS ONE* **7**, e34222. (doi:10.1371/journal.pone.0034222)

87. Carlson J, Doyle J. 1999 Highly optimized tolerance: a mechanism for power laws in designed systems. *Phys. Rev. E* **60**, 1412–1427. (doi:10.1103/PhysRevE.60.1412)

88. Carlson J, Doyle J. 2002 Complexity and robustness. *Proc. Natl Acad. Sci. USA* **99**, 2538–2545. (doi:10.1073/pnas.012582499)

89. Newman M, Girvan M, Farmer J. 2002 Optimal design, robustness, and risk aversion. *Phys. Rev. Lett.* **89**, 28301. (doi:10.1103/PhysRevLett.89.028301)

90. Collins HM. 2001 Tacit knowledge, trust and the Q of sapphire. *Soc. Stud. Sci.* **31**, 71–85. (doi:10.1177/030631201031001004)

91. Beveridge WIB. 1957 *The art of scientific investigation*. New York, NY: W. W. Norton.

92. Sutherland WJ, Burgman MA. 2015 Use experts wisely. *Nature* **526**, 317–318. (doi:10.1038/526317a)

93. Horwitz W, Albert R. 2006 The Horwitz ratio (HorRat): a useful index of method performance with respect to precision. *J. AOAC Int.* **89**, 1095–1109.

94. Eisler Z, Bartos I, Kertesz J. 2008 Fluctuation scaling in complex systems: Taylor's law and beyond. *Adv. Phys.* **57**, 89–142. (doi:10.1080/00018730801893043)

95. Thalmann R. 2002 CCL key comparison: calibration of gauge blocks by interferometry. *Metrologia* **39**, 165–177. (doi:10.1088/0026-1394/39/2/6)

96. Lange KL, Little RJA, Taylor JMG. 1989 Robust statistical modeling using the *t*-distribution. *J. Am. Stat. Assoc.* **84**, 881–896. (doi:10.2307/2290063)

97. Gott III JR, Vogeley MS, Podariu S, Ratra B. 2001 Median statistics, H_0, and the accelerating universe. *Astrophys. J.* **549**, 1–17. (doi:10.1086/319055)

98. Avella Medina M, Ronchetti E. 2015 Robust statistics: a selective overview and new directions. *Wiley Interdiscip. Rev. Comput. Stat.* **7**, 372–393. (doi:10.1002/wics.1363)

99. Gelman A. 2006 Prior distributions for variance parameters in hierarchical models (comment on an article by Browne and Draper). *Bayesian Anal.* **1**, 515–533. (doi:10.1214/06-BA117A)

100. Polson NG, Scott JG. 2012 On the half-Cauchy prior for a global scale parameter. *Bayesian Anal.* **7**, 887–901. (doi:10.1214/12-BA730)

101. Johnson VE. 2013 Revised standards for statistical evidence. *Proc. Natl Acad. Sci. USA* **110**, 19 313–19 317. (doi:10.1073/pnas.1313476110)

102. Gelman A, Robert CP. 2014 Revised evidence for statistical standards. *Proc. Natl Acad. Sci. USA* **111**, E1933. (doi:10.1073/pnas.1322995111)

103. Colquhoun D. 2014 An investigation of the false discovery rate and the misinterpretation of p-values. *R. Soc. open sci.* **1**, 140216. (doi:10.1098/rsos.140216)

104. Fischhoff B, Davis AL. 2014 Communicating scientific uncertainty. *Proc. Natl Acad. Sci. USA* **111**, 13 664–13 671. (doi:10.1073/pnas.1317504111)