Scientific measurements are never perfect, but different teams studying the same phenomena can often produce widely different results. These "outlier" values are the cause of much consternation – but they may also be a sign of healthy scientific progress, says **David Bailey**

Why OUTLIERS are good for science

David Bailey is a professor in the Department of Physics, University of Toronto. He worked for many years in experimental high-energy physics, usually looking for things that were not found, so it was a pleasant change to have been one of the many thousands of co-discoverers of the Higgs boson. His current research and teaching focuses on experimental uncertainty and why measurements are sometimes wrong.

Children are often taught in school that measurement is a straightforward process with little room for error. Hold a ruler up to an object, or place some quantity on a scale, and read off the value. The pencil is 12 cm long; the blocks weigh 1.4 kg. Repeat the measurement and the answer will be the same.

In science, measurement is anything but certain. The tools are more complex than a ruler or scale. The objects of study may be less tangible and more ephemeral than pencils and building blocks, and they range in size from the smallest particles of matter to the largest structures in the known universe. All this complexity means that no measurement is perfect. There is always some error.

Scientists treat this error as uncertainty, reporting both the measured value and a range within which any error would be expected to fall. Yet scientific measurements often disagree by far more than their reported uncertainties. Such differences raise questions about how uncertainties are evaluated and interpreted, whether some outliers are an unavoidable consequence of good science at work, and what are realistic expectations for scientific reproducibility.

Big G

Most recent concern about irreproducible research has focused on life and social sciences, but the scientific quantity with the longest history of inconsistent measurements may be Newton's constant, G_N. "Big G", as it is often called, determines the strength of gravity and is important from the tiniest to the largest scales in physics – from setting the size of superstrings to determining the motions of planets, stars and galaxies. It has been carefully measured by excellent scientists for over two hundred years.

These scientists worried about and investigated every possible source of error they could think of, but how well did they estimate the uncertainties in their results? Since the true value of G_N is unknown (which is why we must measure it), the only way to judge the accuracy of reported uncertainties is to check if the differences between measured values are consistent with their uncertainties. These differences are shown in Figure 1.

One might hope that the reported uncertainties, σ ("sigma"), would be associated with familiar "bell curve" Gaussian probabilities, in which case we would expect only 5% of the differences to be more than 2σ, and less than one in a million to be more than 5σ. The data are not even close, with 35% of the differences more than 2σ, 17% greater than 5σ, and one measurement disagreeing by 65σ! Instead of being Gaussian, the differences are consistent with the Cauchy distribution, infamously so broad that its average value does not even exist, and some statisticians have had difficulty believing that it could be associated with the uncertainty of any sensible measurement.

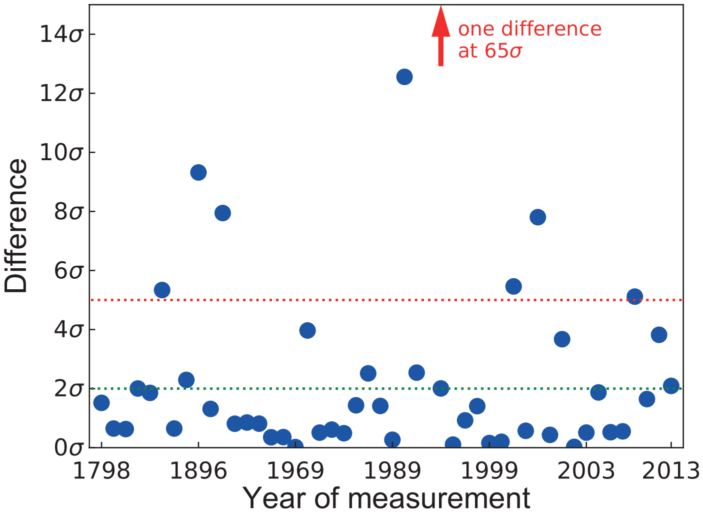


FIGURE 1 Absolute differences between measurements of G_N and its current official CODATA 2014 value of (6.67408 ± 0.00032) × 10^{-11} m^3 kg^{-1} s^{-2}, in units of the estimated uncertainty (σ) for each difference. One difference is 65σ, far off the top of the chart. The horizontal dotted lines are at 2 and 5σ.

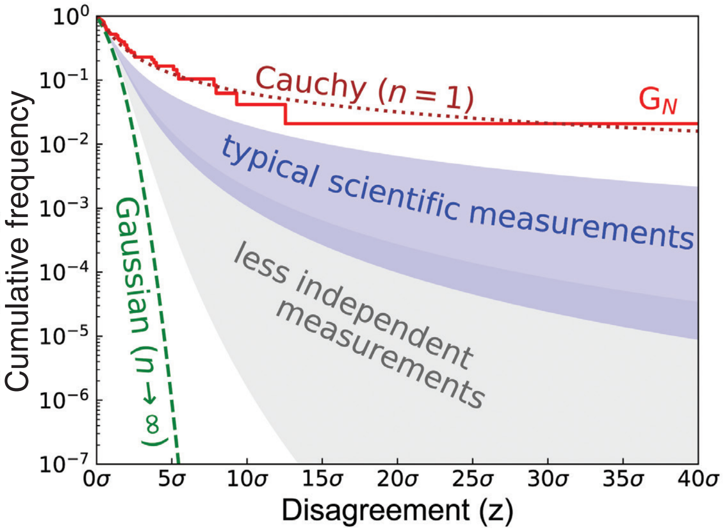


FIGURE 2 Observed cumulative frequencies with which two measurements disagree by more than zσ. The blue area is where independent scientific measurements typically lie; the grey area shows the region for measurements with possible shared biases. Challenging measurements can have even heavier tails: the observed distribution (in red) of measurements of Newton's constant, G_N, is very close to a Cauchy distribution.

Scientific accuracy

The problem is not only with Big G. Although Newton's constant is an extreme example, non-Gaussian outliers are common for scientific measurements. This is illustrated in Figure 2, based on a study of 40 526 measurements of 3228 quantities in fields from medicine to particle physics.[1] The distributions of differences between measurements of the same quantities are well described by Student's *t* distributions (see box, page 16). For Gaussian uncertainties, we would expect to follow the dashed green curve. But what is seen for typical scientific measurements are distributions with long power-law ($1/z^{n+1}$) tails (with n around 2 or 3), far from Gaussian. Instead of less than one in a million, the chance of a greater than 5σ outlier can be more than one in 10, and is rarely less than 1%.

These broad distributions are not just because careless scientists sometimes make mistakes. Even the best researchers never completely understand their measurements, and since the best scientists tend to work at the cutting edge, they are more likely to run into trouble with new and subtle problems. The most discrepant Big G measurement in Figure 1 is a 1996 result from a team at the national metrology institute of Germany. Despite their best efforts at the time, they could find no explanation for the 65σ disagreement. It took 8 years to identify the probable cause: overlooked variations in the electrical properties of an innovative component.

Other leading national metrology labs have had similar experiences. The US National Institute for Standards and Technology (NIST) reported many nuclear lifetimes that differed from results from other labs, until it was realised that a sample-positioning ring in the NIST equipment had been slowly slipping over several decades of use. A 9σ discrepant 1999 value of Avogadro's number by Japan's national measurement lab, using a new method, was traced back to imperfections in the silicon crystals being studied.

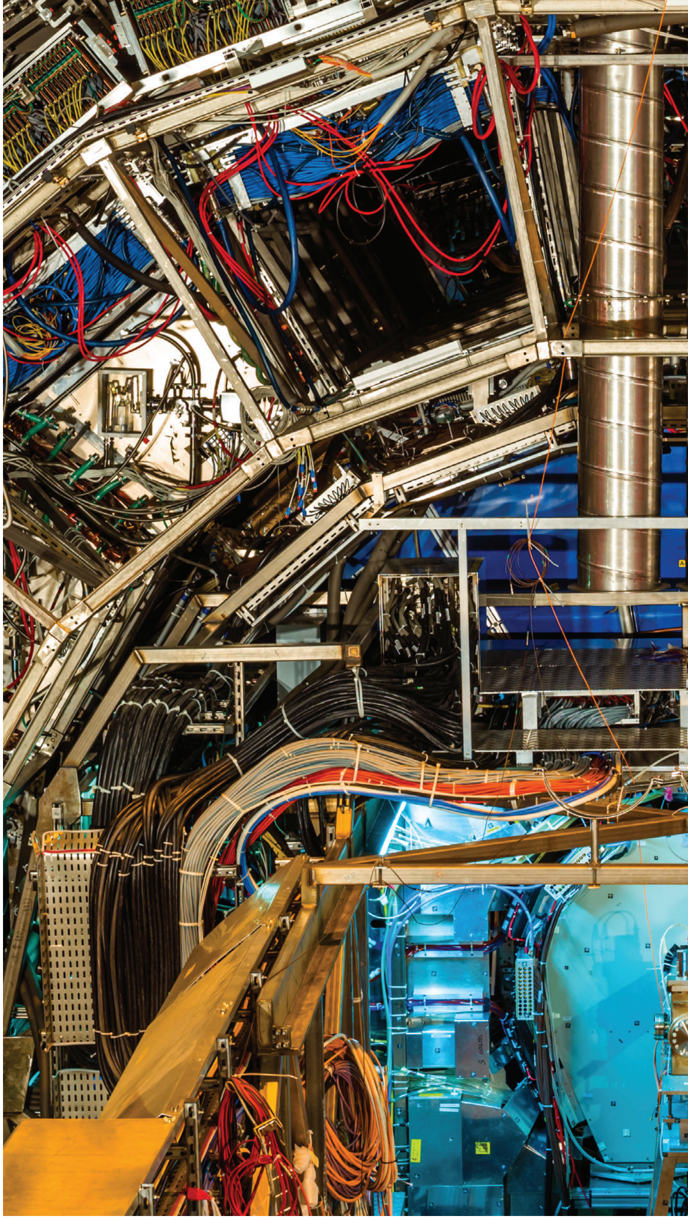
New ideas produce new problems, so outliers can be the result of creativity instead of incompetence. The more independent a measurement is – with differences in people, methods, location, times, apparatus, software, or models – the more likely it is to disagree with other results. Less independent measurements, such as those using the same type of instrument under similar conditions, are usually much more Gaussian, with n as large as 10 and the chance of a 5σ outlier being lower than 0.1%. Such measurements have a high probability of sharing biases, so are likely to agree even when they are wrong. The best way to find such errors is to make measurements with different methods and improved precision.

Complex uncertainty

Every measurement comes with uncertainties from many sources. Scientists love statistical uncertainties due to random independent variations. Such uncertainties can be evaluated using well-established methods with strong theoretical foundations, and they become smaller and usually more Gaussian with repeated measurements.

Unfortunately, much uncertainty is associated with possible sources of bias in the complex systems associated with any scientific measurement. These uncertainties do not automatically become smaller with more data, and evaluating them is sometimes more art than science. Bias can occur in any aspect of a measurement, including wrong calibrations, non-random samples, misunderstood backgrounds, or inadequate theory.

Uncertainties associated with systematic errors are estimated using models that include everything known about



iStock/xenotar

Student's *t* distributions

The generalised Student's t distribution with n degrees of freedom

$$S_{n,\sigma}(z)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}\sigma}\left(1+\frac{1}{n}\left(\frac{z}{\sigma}\right)^2\right)^{-\frac{n+1}{2}}$$

is essentially a smoothly symmetric normalisable power law, with

$S_{n,\sigma}(z)\sim 1/(z/\sigma)^{n+1}$ for $|z|\gg\sqrt{n}\sigma$

The parameter σ defines the core width and overall scale of the distribution, and is equal to the standard deviation in the $n\to\infty$ limit where the t distribution converges to the Gaussian distribution:

$$N_\sigma(z)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}$$

The parameter n determines the size of the tail, with small n corresponding to large tails; n = 1 is the Cauchy distribution where, roughly speaking, half the distribution is in the heavy tails.

iStock/easyturn

ABOVE Either what is measured or how it is measured can be complex: crowds of people or the ALICE detector at CERN.

the measurement, but unknown systematic errors can produce large non-Gaussian outliers. As David Hand puts it: "Things which ought to be expected can seem quite extraordinary if you've got the wrong model".[2]

Some outliers appear to be unavoidable consequences of Murphy's law – "That which can go wrong, will go wrong" – acting on any complex system where many things can go wrong with consequences of all sizes. Failures in designed complex systems – from software to electrical grids to nuclear reactors – often have power-law distributions,[1] so the observed heavy tails in scientific consistency should be no surprise.

These frequent errors are not due to any lack of effort to avoid them. Good scientists have "experimental paranoia" – the strong belief that the universe is conspiring to ruin their measurement. They take great care to verify and validate everything, and search for inconsistencies in their understanding.

The best way to evaluate known systematic effects and detect unknown errors is to stress-test the measurement model by changing something in the apparatus, procedures or analysis, and seeing if the new result is consistent with the model. Scientists change everything they can think of – time, people, samples, apparatus, temperature, software, and more – seeing how the system responds and constantly looking for unexpected results. The changes are made as extreme as possible without (usually) breaking anything too expensive. The uncertainty due to known systematic effects is evaluated based on the observed variations. The causes of any inconsistencies are investigated, and any systematic errors found are either fixed or their effect included in the reported uncertainty.

For example, measuring Big G is subject to myriad subtle problems. The gravitational force between known test masses must be measured, and this force is so small that it can be overwhelmed by a breath of air or a few stray electric charges. When Henry Cavendish made the first measurements of Big G in 1798, he noticed that the attraction between the masses slowly drifted during each measurement. Magnetism was not well understood at the time, so he wondered whether his test masses were being slowly polarised by the Earth's magnetic field. He replaced some of the masses with actual magnets, and when no change in his result was seen he was confident that magnetism was not the problem. He had thought that temperature differences in his apparatus would be too small to generate problematic air currents, but to be sure he placed a thermometer in his lab and noticed that there was a small temperature increase over the duration of each measurement. He then added more thermometers and studied the temperature sensitivity by heating the masses with oil lamps and then cooling them with ice. This showed that the drift was associated with temperature differences between the masses and the surrounding case that were producing otherwise imperceptible air currents. Foreshadowing modern detector modelling, Cavendish also calculated the gravitational force exerted by the experiment's mahogany case on the test masses to make sure this was not large enough to perturb his results. He took measurements in all weathers and with many other changes in his apparatus, and his final result included an uncertainty based on all the variations observed.

Erratic systematic effects can be particularly tricky to understand. Variations in a measurement of the Z boson mass at CERN, the European centre for particle physics, were initially larger than desired. The physicists had already corrected for subtle effects, such as the deformation of the particle accelerator due to monthly Earth tides and seasonal changes in local water levels, but only after additional monitoring did they notice fluctuations in the particle beam energy that were eventually matched with the local train schedule. Electrical currents from nearby railway lines, such as the Geneva–Paris TGV, were passing through the accelerator and changing the beam energy. Once understood, these effects could be corrected and the best mass resolution achieved.

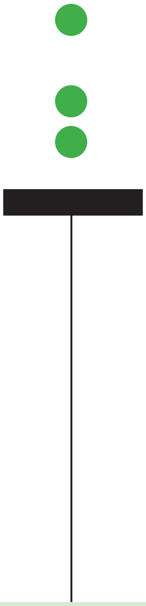
Even when an inconsistency is noticed, however, it cannot always be understood. The reason for a discordant NIST value of Planck's constant was never found despite the lab bringing in a completely new team to go over every component and procedure. Later measurements by the same group were not anomalous; the unknown problem had simply gone away – a disquieting but familiar experience for many scientists.

Scientists try, of course, to design their measurements to be tolerant of known uncertainties and risks, but this can actually make them more sensitive to unknown problems.[3] For example, you can reduce the temperature sensitivity of your apparatus by putting it inside an insulated box, but if that box vibrates because it resonates with the previously unnoticeable hum from your air conditioner, your shaken measurements may end up being much worse instead of better.

Precise, accurate, affordable – pick two

No amount of effort can ever prove that a measurement is free of every source of error, and no-one has infinite money or time, so choices must be made. Scientists must decide how to allocate resources between improving statistical precision (by taking more primary data) and reducing the chance of being wrong (by searching for systematic errors with more consistency checks).

Although better measurement precision improves understanding of systematic effects, it does not automatically reduce the heaviness of the uncertainty tails. Improved precision allows the signal to be examined and understood in more detail, possibly revealing undetected systematic errors, but it also means that unknown errors that were previously too small to matter can now become consequential. For every systematic effect newly understood because of better precision, a new unknown systematic error is likely to become important, leaving the shape of the uncertainty tail unchanged.



Rapidly improving precision can even lead to heavier tails, since there is less time to understand existing methods before they are replaced by more precise methods with new errors. This may be why uncertainties in physics often have heavier tails than in medical research.[1] Constantly improving technology means that a new physics measurement is typically twice as precise as the best previous measurement of the same quantity. Improving precision in medicine can be expensive since halving the sampling error will require studying four times as many people at substantially greater cost, so new studies in medical research are frequently less precise than the best previous similar measurements.

The electron charge

In 1913, University of Chicago physicist Robert Millikan published studies of charged oil drops, which showed that electric charge was quantised and determined the charge on the electron (*e*) with 15 times more precision than previous measurements. Over the next decade, Millikan and his current or former students published further results that agreed with this measurement, and Millikan was awarded the Nobel Prize in Physics in 1923. The first significant measurement by a different group did not come until 1928, after a new X-ray method for determining *e* was developed. Although the first new measurement (by a Chicago graduate student) agreed with Millikan's value for *e*, subsequent measurements by other researchers did not. Either the oil-drop or X-ray method had an unknown bias, and both were subject to much scrutiny. By 1936 it was apparent that Millikan had used a slightly inaccurate viscosity of air, and with a new, more accurate viscosity value, the old oil-drop measurements were found to be consistent with the new X-ray results. By 1938 even Millikan agreed that his earlier values were wrong. A series of beautifully consistent results had shared a common bias that only became apparent when experiments were done by different groups using different methods.

Better precision also allows physicists to set tough "significance" criteria. Any discovery claim in particle physics is expected to be based on at least a 5σ effect, and although this convention evolved simply from a desire to keep the rate of false discoveries down to a manageable level, it also ensures that there is sufficient precision for internal consistency checks. It is hard to make statistically significant comparisons of subsets of a 2σ signal, but a 5σ signal can be sliced in many ways and checked for consistency. For example, physicists were confident in the discovery of the Higgs boson not just because the total signal was greater that 5σ, but because it was observed in two experiments, in data taken at different times and energies, and in multiple detection channels. Of course, even a 5σ "significant" signal can be wrong, as illustrated in recent years by 6σ reports of faster-than-light neutrinos and cosmic inflation that turned out to be the result of bad cables and background galactic dust.

Diversity matters

Although outliers are usually considered bad, the absence of outliers can sometimes be worse. As noted long ago by Harold Jeffreys,[4] if differences between nominally independent measurements appear to be Gaussian, it may just mean that they share most of their systematic errors, not that they have none. Different researchers using diverse and constantly improving methods should, when wrong, be wrong in different ways and disagree (see "The electron charge").

As much as scientists would like to believe their experiments and models are accurate, all complex systems are subject to failure and some outliers are unavoidable. Scientists are generally good at estimating typical uncertainties, but are often hesitant to acknowledge the long tails of uncertainty that produce outliers. This reluctance is in part because, unlike the typical uncertainty σ, the exponent *n* of the uncertainty tail for an individual result cannot be evaluated by conventional methods. Instead of ignoring what we cannot calculate, however, it is better to use past history to predict how likely it is that a new measurement is wrong, and to accept that some outliers are an inevitable consequence of healthy scientific progress. ■

References

1. Bailey, D. (2017) Not Normal: the uncertainties of scientific measurements. *Royal Society Open Science*, **4**, 160600.

2. Watkins, P. and Hand, D. (2014) Things which ought to be expected can seem quite extraordinary if you've got the wrong model. *Significance*, **11**(4), 36–39.

3. Carlson, J. M. and Doyle, J. (1999) Highly optimized tolerance: A mechanism for power laws in designed systems. *Physical Review E*, **60** (2A), 1412–1427.

4. Jeffreys, H. (1938) The law of error and the combination of observations. *Philosophical Transactions of the Royal Society of London*, *Series A*, **237**, 231–271.

SDSS
SYMPOSIUM ON DATA SCIENCE & STATISTICS

Analytics, Computational Statistics, and Visualization

Sessions will be centered on the following topic areas:

- STATISTICAL MACHINE LEARNING
- DATA VISUALIZATION
- DATA SCIENCE
- APPLICATIONS
- COMPUTATIONAL STATISTICS
- COMPUTING SCIENCE

ATTEND

Online Registration and Housing Open: February 1

Speaker Registration Deadline: March 15

Early Registration Deadline: April 5

Housing Reservation Deadline: April 23

Online Registration Deadline: May 15

Onsite Registration: May 16–19

Learn more at ***ww2.amstat.org/sdss***.

BEYOND BIG DATA: LEADING THE WAY

RESTON, VIRGINIA • MAY 16–19, 2018

ASA