

Pay No Attention to the Model Behind the Curtain

Philip B. STARK

Department of Statistics

University of California

Berkeley, CA 94720-3860

www.stat.berkeley.edu/~stark

To appear in *Significant Digits: Responsible use of quantitative Information*,
Andrea Saltelli and Ângela Guimarães Pereira Eds.

Version: 13 December 2017

The title, ‘Pay No Attention to the Model Behind the Curtain,’ is a reference to the 1939 film *The Wizard of Oz*. When Dorothy, the Tin Man, the Cowardly Lion, and the Scarecrow first meet the Wizard, he intimidates them, manifesting as a booming voice and a disembodied head amid towering flames, steam, smoke, and colored lights. During this audience, Dorothy’s dog Toto tugs back a curtain, revealing the Wizard to be a small, ordinary man at the controls of a big machine that produces impressive effects to distract and impress onlookers. The Wizard’s artificially amplified, booming voice instructs Dorothy and her entourage to pay no attention to the man behind the curtain.

The little man behind the curtain is an apt metaphor for the role of statistical models in science and policy. In a vast number of cases—most, perhaps—impressive computer results and quantitative statements about probability, risk, expected costs, and other putatively solid facts are actually controlled by a model behind the curtain, a model that nobody’s supposed to look at carefully or pay attention to. What appears to be impressive ‘science’ is in fact an artificial amplification of the opinions and *ad hoc* choices built into the model, which has a heuristic basis rather than a tested (or even testable) scientific basis. The levers in the model—parametrizations, transformations, assumptions about the generative mechanism for the data, data uncertainties, and such—are like the levers in the Wizard’s machine. And just like the Wizard’s machine, the models have the power to persuade and intimidate, but not the power to predict or to control.

Quantifauxcation.

Relying on ungrounded, untested, *ad hoc* statistical models is *quantifauxcation*, a neologism for the process of assigning a meaningless number, then pretending that because the result is quantitative, it must mean something (and if the number has six digits of precision, they all matter). Quantifauxcation usually involves some combination of data, pure invention, invented models, inappropriate use of statistics, and logical lacunae. It is often involved in ‘informing’ policy.

Cost-benefit analyses are an example. It’s widely claimed that the only rational basis for policy is a quantitative cost-benefit analysis.¹ But if there’s no rational basis for its quantitative inputs, how can a cost-benefit analysis be a rational basis for anything? Not only are costs and consequences hard to anticipate, enumerate, or estimate in real-world problems, but behind

¹ https://en.wikipedia.org/wiki/Cost%E2%80%93benefit_analysis

every cost-benefit analysis is the tacit assumption that all costs and all benefits can be put on a common, linear scale, such as money or ‘utility.’

The idea that you can convert outcomes into dollars, ‘utils,’ or ‘quality-adjusted life years,’ to put them on the same scale is an *assumption*.² It’s neither obviously true nor even necessarily true. As a matter of mathematics, multidimensional spaces—such as competing objectives—are not in general totally ordered. For example, to put all consequences on a monetary scale, you have to assign a dollar value to human life, including future generations; to environmental degradation; to human culture; to endangered species; and so on. Some people are reluctant to be so draconian.³

Similarly, there’s a slogan that *risk equals probability times consequences*. But what if probability doesn’t apply to the phenomenon you’re talking about? What if you can’t quantify the consequences, or can’t put them on a common scale? Insisting on quantifying risk and on quantitative cost-benefit analyses requires doing things that may not make sense technically or morally. Moreover, I’ve yet to see a compelling example of incorporating uncertainty in the estimate of the probability (if the notion of ‘probability’ applies to the problem at all) and uncertainty in the consequences—much less the ‘value’ of those consequences.

Theories of Probability.

What is probability? It has an axiomatic aspect and a philosophical aspect. Kolmogorov’s axioms, the mathematical basis of modern probability, are just that: math. *Theories of probability* provide the glue to connect the mathematics to the real world, allowing us to interpret probability statements.⁴

The oldest interpretation of probability, *equally likely outcomes*, arose from studying games of chance, in particular, dice games. This theory argues that if the system is symmetric, like a well balanced die, there’s no reason Nature should prefer one outcome to another, so all outcomes are equally likely. The theory argues that if a die is symmetric and balanced, there’s no reason for it to land with any particular side on top. Therefore, all six possible outcomes are equally likely, from which many consequences follow as a matter of mathematics.

This interpretation has trouble in situations that don’t have the intrinsic symmetries coins, dice, and roulette wheels have. For example, suppose that instead of rolling a die, you’re tossing a thumbtack. What’s the probability that it lands point up versus point down? There’s no obvious symmetry to exploit. Should you simply deem those two outcomes to be equally likely? And how might you use symmetry to make sense of ‘the probability of an act of nuclear terrorism in the year 2016?’ That’s even weirder. There many events for which it’s impossible to define ‘probability’ using equally likely outcomes.

² See, e.g., Luce, R.D., and J.W. Tukey, 1964. Simultaneous conjoint measurement: A new type of fundamental measurement, *Journal of Mathematical Psychology*, 1, 1–27.

³ See, e.g., Funtowicz, S.O., and Ravetz, J.R., 1994. The worth of a songbird: ecological economics as a post-normal science. *Ecological Economics*, 10, 197–207.

⁴ For a more technical discussion, see Freedman, D.A., 2010. Issues in the Foundations of Statistics: Probability and Statistical Models, and P.B. Stark and D.A. Freedman, 2010, What is the Chance of an Earthquake, both in *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, D. Collier, J. Sekhon, and P.B. Stark, eds., Cambridge University Press, NY. For an elementary discussion, see Stark, P.B., 1997. *SticiGui*, Chapter 13 Probability: Philosophy and Mathematical Background, <http://www.stat.berkeley.edu/~stark/SticiGui/Text/probabilityPhilosophy.htm>

The second approach is the *frequency theory*, which defines probability in terms of limiting relative frequencies. According to the frequency theory, what it means to say ‘the chance that a coin lands heads’ is that if one were to toss the coin again and again, the fraction of tosses that resulted in heads would converge to a limit; that limit is *defined* to be the probability of heads.

There are many phenomena for which this approach to defining probability makes sense and many for which it doesn’t, for instance, if, as a matter of principle, the phenomenon cannot be repeated. What’s the probability that global average temperature will increase by three degrees in the next 50 years? Can we in principle repeat the next 50 years over and over and over again to see what fraction of the time that happens?

The *subjective theory* or *Bayesian theory* may work in such situations. It defines probability in terms of degree of belief. According to the subjective theory, what it means to say “the probability that a coin lands heads is 50%” is that the speaker believes with equal strength that it will land heads as he or she believes that it will land tails. Probability thus measures the state of mind of the person making the probability statement. That has a number of problems, among them, why should I care what your internal state of mind is? Compared with both the theory of equally likely outcomes and the frequency theory, the subjective theory changes the subject. The theory of equally likely outcomes is about the symmetry of the *coin*. The frequency theory is about what the *coin* will do in repeated tosses. The subjective theory is about what *I think*. It changes the subject from geometry or physics to psychology. The situation is complicated further by the fact that people are actually not very good judges of what’s going to happen, as discussed below. For making personal decisions, for instance, deciding what to bet one’s own money on, the subjective theory may be an excellent choice.

Another approach, different from these three classical interpretations, is to treat probability models as *empirical commitments*.⁵ For instance, coin tosses are not random. If you knew exactly the mass distribution of the coin, its initial angular velocity, and its initial linear velocity, you could predict with certainty how a coin would land. But you might prefer to model the toss as random, at least for some purposes. Modelling it as random entails predictions that can be checked empirically against data. The usual model of tosses as fair and independent implies that all 2^n possible sequences of heads and tails in n tosses of a fair coin are equally likely, implying probability distributions for the lengths of runs, the number of heads, etc. If you compare this model to data, you will find that if the number of tosses is sufficiently large, the model does not fit accurately. There tends to be serial correlation among the tosses. And the frequency of heads will tend to be ‘surprisingly’ far from 50%.⁶

The last interpretation of probability is probability as *metaphor*. This seems to be how probability enters most policy applications. It does not assert that the world truly behaves in a random way. Rather, it says that a phenomenon occurs ‘as if’ it were a casino game. This is closely tied to probability as an empirical commitment, although the step of checking whether the model matches data is often omitted. We will see more on this below.

⁵ Freedman, D.A., and R.A. Berk, 2010. *Statistical Models as Empirical Commitments*, in *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, D. Collier, J. Sekhon, and P.B. Stark, eds., Cambridge University Press, NY.

⁶ While imprisoned by the Nazis in World War II, John E. Kerrich tossed a coin 10,000 times; it landed heads 5,067 times. https://en.wikipedia.org/wiki/John_Edmund_Kerrich

Probability models in Science.

How does probability enter a scientific problem? It could be that the underlying physical phenomenon really is random, for instance, radioactive decay and other quantum processes. Or it could be that the scientist deliberately introduces randomness, e.g., by conducting a randomized experiment or by drawing a random sample.

Probability can enter as a *subjective prior probability*. Suppose we want to estimate the probability p that a particular coin lands heads (on the assumption that the outcome is random). Surely the probability p is between zero and one. A common subjective approach to capturing that constraint involves positing a *prior probability distribution* for p , for instance, by assuming that p itself was selected at random from the interval $[0, 1]$, according to a uniform probability distribution. Unfortunately, positing any particular probability distribution for p adds an infinite amount of information about the coin, information not contained in the constraint. A constraint is not equivalent to a probability distribution.⁷ A probability distribution is a function, which requires infinitely many numbers to specify (its value at every point in the domain). The function—the prior—needs to be selected from an infinite-dimensional set of possibilities.

Theoretical results say that under some conditions, the prior does not matter asymptotically: the data ‘swamp’ the prior. Those results involve conditions that might not hold in practice. In particular, it isn’t true in general for infinite-dimensional unknowns or for improper priors.⁸ Nor is it necessarily true if the dimensionality of the problem grows as the number of data grows.⁹ Nor is it true that nominally ‘uninformative’ (i.e., uniform) priors are actually uninformative, especially in high-dimensional spaces.¹⁰

Beyond the technical difficulties, there are practical issues in eliciting prior distributions, even in one-dimensional problems.¹¹ In fact, priors are almost never elicited—rather, priors are chosen for mathematical convenience or from habit. There are arguments in favor of the Bayesian approach from *Dutch book*: if you are forced to cover all possible bets and you do

⁷ See Stark, 2015. Constraints versus priors, *SIAM/ASA Journal of Uncertainty Quantification*, 3, 586–598. doi:10.1137/130920721

<http://epubs.siam.org/doi/10.1137/130920721>; Stark, P.B., and Tenorio, L., 2010. A Primer of Frequentist and Bayesian Inference in Inverse Problems, in *Large Scale Inverse Problems and Quantification of Uncertainty*, Biegler, L., G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders and K. Willcox, eds. John Wiley and Sons, NY.

⁸ See, e.g., Freedman, D.A., 1999. Wald Lecture: On the Bernstein-von Mises Theorem with Infinite Dimensional Parameters, *Annals of Statistics*, 27, 1119–1141.

⁹ Diaconis, P. and D.A. Freedman, 1986. On the Consistency of Bayes Estimates, *Annals of Statistics*, 14, 1–26.

¹⁰ Stark, 2015. Constraints versus priors, *SIAM/ASA Journal of Uncertainty Quantification*, 3, 586–598. doi:10.1137/130920721 <http://epubs.siam.org/doi/10.1137/130920721>; Backus, G.E., 1987. Isotropic probability measures in infinite-dimensional spaces, *Proceedings of the National Academy of Science*, 84, 8755–8757.

¹¹ See, e.g., O’Hagan, A., 1998. Eliciting Expert Beliefs in Substantial Practical Applications, *Journal of the Royal Statistical Society. Series D (the Statistician)* 47, 21–35.

<http://www.jstor.org/stable/2988425>. Typically, some functional form is posited for the distribution, and only some parameters of that distribution, such as the mean and variance or a few percentiles, are elicited.

not bet according to a Bayesian prior, there are collections of bets where you are guaranteed to lose money, no matter what happens. According to the argument, you are therefore not rational if you don't bet in a Bayesian way. But of course, one is not forced to place bets on all possible outcomes.¹²

A fourth way probability can enter a scientific problem is through the invention of a *probability model* that's supposed to describe a phenomenon. But in what sense is the model supposed to describe the phenomenon, to what level of accuracy, and for what purpose? Describing data tersely (a sort of data compression), predicting what a system will do next, and predicting what a system will do in response to a particular intervention are very different goals. The last involves causal inference, which is far more difficult than the first two.¹³ Creating a model that has a parameter called 'probability' and fitting the model to data does not mean that the estimated value of the parameter is in fact the probability of anything. Just as the map is not the territory, the model is not the phenomenon, and calling something 'probability' does not make it so.

Finally, probability can enter a scientific problem as metaphor: a claim that the phenomenon in question behaves 'as if' it is random. What 'as if' means is rarely made precise, but this approach is common, for instance, in stochastic models for seismicity.¹⁴

Creating randomness by taking a random sample or assigning subjects at random to experimental conditions is quite different from *inventing* a probability model or proposing a metaphor. The first may allow inferences if the analysis properly takes into account the randomization, if there are adequate controls, and if the study population adequately matches the population for which inferences are sought. But when the probability exists only within an invented model or as a metaphor, the inferences have little foundation. They are no better than the assumptions. The assumptions and the sensitivity of the conclusions to violations of the assumptions have to be checked in each application and for each set of data.

In summary, the word 'probability' is often used with little thought about why, if at all, the term applies, and many standard uses of the word are rather removed from anything actually random.

Uncertainty versus Probability.

Many scientists also instinctively use the word 'probability' to describe anything uncertain. But not all uncertainties can be represented as probabilities. A common taxonomy divides uncertainties into *aleatory* and *epistemic* uncertainties. Aleatory uncertainty describes randomness arising from the play of chance mechanisms—the luck of the draw. Epistemic

¹² See Freedman, D.A., 2010. Issues in the Foundations of Statistics: Probability and Statistical Models, in *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, D. Collier, J. Sekhon, and P.B. Stark, eds., Cambridge University Press, NY.

¹³ See Freedman, D.A., 2010. The Grand Leap, in *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, D. Collier, J. Sekhon, and P.B. Stark, eds., Cambridge University Press, NY.

¹⁴ See, e.g., Stein, S., and Stein, J., 2013. Shallow versus deep uncertainties in natural hazard assessments, *EOS*, 94, 133–140.

<http://onlinelibrary.wiley.com/doi/10.1002/2013EO140001/epdf>

This paper says the occurrence of earthquakes and other natural hazards is like drawing balls from an urn, and makes distinctions according to whether and how the number of balls of each type changes between draws. But why is the occurrence of natural hazards like drawing balls from an urn? This metaphor has no basis in physics.

uncertainty results from ignorance, rather than chance. Epistemic uncertainty is ‘stuff we don’t know’ but in principle could learn.

Canonical examples of aleatory uncertainty include coin tosses, die rolls, lotteries, radioactive decay, some kinds of measurement error, and the like. Under some circumstances, such things do behave (approximately) as if random—but generally not perfectly so, as mentioned above. Canonical examples of epistemic uncertainty include ignorance of the physical laws that govern a system or ignorance of the values of parameters in a system.

Imagine a biased coin that has an unknown chance p of landing heads. Ignorance of the chance of heads is epistemic uncertainty. But even if we knew the chance of heads, we would not know the outcome of the next toss: it would still have aleatory uncertainty.

Expert Opinion and Prior Probabilities.

The standard way to combine aleatory and epistemic uncertainties involves using subjective (aka *Bayesian*) prior probability to represent epistemic uncertainty. In effect, this puts individual beliefs on a par with an unbiased physical measurement that has a known uncertainty. We do know that the chance of heads must be between 0 and 1, but we do not know more than that. Attempting to combine aleatory and epistemic uncertainties by representing them both as probabilities amounts to saying that I could weigh some object on an actual physical scale or I could think hard about how much it weighs: the two are on a par. Thinking hard about the question produces an unbiased measurement that has an uncertainty, just like the scale has an uncertainty. Moreover, I know the accuracy of my internal ‘measurement’ from careful introspection. Hence, I can combine the two sources of uncertainty as if they are independent measurements of the same thing, both made by unbiased instruments.¹⁵

This doesn’t work.

Sir Francis Bacon’s triumph over Aristotle should have put this idea to rest: in general, it is not possible to make sound inferences about the world by just ratiocination (some *Gedankenexperiments* notwithstanding¹⁶). Psychology, psychophysics, and psychometrics have shown empirically that people are bad at making even rough qualitative estimates, and that quantitative estimates are usually biased.

Moreover, the bias can be manipulated through processes such as *anchoring* and *priming*, as described in the seminal work of Tversky and Kahneman.¹⁷ Anchoring doesn’t just affect individuals—it affects entire disciplines. The Millikan oil drop experiment¹⁸ to measure the

¹⁵ When practitioners analyse complex systems such as climate, the economy, earthquakes, and such, the same observations they use as data in the problem are also the basis of their beliefs as reflected in the prior. But the analysis generally treats the data and prior as if they provided “independent” measurements—another fishy aspect of this approach.

¹⁶ A favorite example is Galileo’s ‘demonstration’ that Aristotle was wrong that bodies of different masses fall at different rates: evidently, Galileo refuted Aristotle using a thought experiment. https://en.wikipedia.org/wiki/Galileo%27s_Leaning_Tower_of_Pisa_experiment (last accessed 16 November 2016)

¹⁷ Tversky, A., and Kahneman, D., 1975. Judgment under uncertainty: heuristics and biases. *Science*, 185, 1124–1131.

¹⁸ Millikan, R.A., 1913. On the Elementary Electrical Charge and the Avogadro Constant, *Physical Review*, 2, 109–143. <http://history.aip.org/history/exhibits/gap/PDF/millikan.pdf> (last accessed 16 November 2016)

charge of an electron is an example: Millikan's value was too low, supposedly because he used an incorrect value for the viscosity of air. It took about 60 years for new estimates to "drift up" towards the currently accepted value, which is about 0.8% higher (a small difference, but considerably larger than the error bars). Other examples include measurements of the speed of light and beliefs about the amount of iron in spinach.¹⁹ In these examples and others, somebody got something wrong and it took a very long time for a discipline to correct the error because subsequent exponents did not stray too far from the previous estimate, perhaps because the existence of the first estimate made them doubt results that were far from it.

Tversky and Kahneman also showed that people are poor judges of probability, subject to strong biases from anchoring and from *representativeness* and *availability*, which in turn depends on the retrievability of instances. Their work also shows that probability judgments are insensitive to prior probabilities and to predictability, and that people ignore the regression to the mean effect—even people who have been trained in probability.

People cannot even accurately judge how much an object weighs with the object in their hands. The direct physical tactile measurement is biased by the density and shape of the object—and even its color.²⁰ The notion that one could just think hard about how global temperature will change in the next 50 years and thereby come up with a meaningful estimate and uncertainty for that estimate is preposterous. Wrapping the estimate with computer simulations barely grounded in physics distracts, rather than illuminates.

Humans are also bad at judging and creating randomness: we have *apophenia* and *pareidolia*, a tendency to see patterns in randomness.²¹ And when we deliberately try to create randomness, what we create has too much regularity. For instance, we avoid 'runs' and repeats.²² People are over-confident about their estimates and their predictions.²³ People's confidence is unrelated to their actual accuracy.²⁴

¹⁹ It's widely believed that spinach has substantially more iron than other green vegetables. It turns out that this is the result of a decimal place error in the 1870s (see, e.g., <http://www.dailymail.co.uk/sciencetech/article-2354580/Popeyes-legendary-love-spinach-actually-misplaced-decimal-point.html>). The published value was off by a factor of 10, because of a transcription error. This has resulted in the common belief that spinach is exceptionally high in iron. The fact that the value was in error was well known before the Popeye character became popular in the 1930s.

²⁰ E.g., Bicchi, A., Buss, M., Ernst, M.O., Peer, A., 2008. *The Sense of Touch and Its Rendering: Progress in Haptics Research*, Springer-Verlag, Berlin Heidelberg (see section 4.4.3)

²¹ <https://en.wikipedia.org/wiki/Apophenia>, <https://en.wikipedia.org/wiki/Pareidolia> (last accessed 16 November 2016)

²² E.g., Schulz, M.-A., Schmalbach, B., Brugger, P., and Witt, K. 2012. Analyzing humanly generated random number sequences: a pattern-based approach. *PLoS ONE* 7, e41531, doi:10.1371; Shermer, M., 2008. Patternicity: Finding Meaningful Patterns in Meaningless Noise, *Scientific American*

²³ E.g., Kahnemann, D., 2011. *Thinking, Fast and Slow*, Farrar, Strauss, and Giroux, NY; Taleb, N.N., 2007. *The Black Swan: The Impact of the Highly Improbable*, Random House, NY.

²⁴ E.g., https://en.wikipedia.org/wiki/Overconfidence_effect (last accessed 16 November 2016); Krug, K., 2007. The Relationship Between Confidence And Accuracy: Current Thoughts of the Literature and a New Area of Research, *Applied Psychology in Criminal Justice*, 2007, 3, 7–41. Chua, E.F., Rand-Giovannetti, E., Schacter, D.L., Albert, M.S., and

If I don't trust your internal scale or your assessment of its accuracy, why should your subjective (Bayesian) analysis carry any weight for me?²⁵

Rates versus probabilities.

It is very common to conflate rates with probabilities. I've seen many examples in the literature of physics, medicine, and other fields.

Evidently, it is a problem in hydrology, too:

The automatic identification of past frequencies with present probabilities is the greatest plague of contemporary statistical and stochastic hydrology. It has become so deeply engrained that it prevents hydrologists from seeing the fundamental difference between the two concepts. It is often difficult to put across the fact that whereas a histogram of frequencies for given quantities [] can be constructed for any function whether it has been generated by deterministic or random mechanism, it can be interpreted as a probability distribution only in the latter case. [] Ergo, automatically to interpret past frequencies as present probabilities means *a priori* to deny the possibility of any signal in the geophysical history; this certainly is not science but sterile scholasticism.

The point then arises, why are these unreasonable assumptions made if it is obvious that probabilistic statements based on them may be grossly misleading, especially when they relate to physically extreme conditions where errors can have catastrophic consequences? The answer seems to be that they provide the only conceptual framework that makes it possible to make probabilistic statements, i.e. they must be used if the objective is to make such probabilistic statements.²⁶

My experience in other branches of physical science and engineering is the same: equating historical rates with probabilities is so routine that it's impossible to get many practitioners to see that there's a profound difference between the two concepts.

Any finite series of dichotomous trials has an empirical rate of success. But the outcomes of a series of trials cannot tell you whether the trials were random in the first place. Suppose there is a series of Bernoulli trials,²⁷ that each trial has the same probability p of success, and that the trials are independent. Then the Law of Large Numbers guarantees that the rate of successes converges (in probability) to the probability of success.

If a sequence of trials *is* random and the chance of success is the same in each trial, then the empirical rate of success is an unbiased estimate of the underlying chance of success. If the trials are random *and* they have the same chance of success *and* you know the dependence structure of the trials (for example, if the trials are independent), then you can quantify the

Sperling, R.A., Dissociating Confidence and Accuracy: Functional Magnetic Resonance Imaging Shows Origins of the Subjective Memory Experience, *Journal of Cognitive Neuroscience* 16, 1131–1142.

²⁵ See Stark, P.B., and Tenorio, L., 2010. A Primer of Frequentist and Bayesian Inference in Inverse Problems, in *Large Scale Inverse Problems and Quantification of Uncertainty*, Biegler, L., G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders and K. Willcox, eds. John Wiley and Sons, NY.

²⁶ Klemeš, V., 1989. The Improbable Probabilities of Extreme Floods and Droughts, in *Hydrology of Disasters: Proceedings of the World Meteorological Organization*, Starosolszky, O., and O.M. Melder, eds., Routledge.

²⁷ Bernoulli trials are random dichotomous trials, each of which can result in *failure* or *success*.

uncertainty of that estimate of the underlying chance. But the mere fact that something has a rate does not mean that it is the result of any random process.

For example, suppose a sequence of heads and tails results from a series of random, independent tosses of a fair coin. Then the rate of heads will converge (in probability) to one half. But suppose I give you the sequence ‘heads, tails, heads, tails, heads, tails, heads, tails, heads, tails, ...’ *ad infinitum*. The limiting rate of heads is $\frac{1}{2}$. Do you think the sequence is random, with 50% chance of heads? Rates are not necessarily the result of anything random, and not necessarily estimates of probabilities.

Here are two thought experiments:

1. You are in a group of 100 people. You are told that one person in the group will die next year. What’s the chance it’s you?
2. You are in a group of 100 people. You are told that one of them is named Philip. What’s the chance it’s you?

There’s not much difference between these two scenarios: both involve a rate of 1% in a group. But in the first one you are invited to say ‘the chance is 1%,’ while in the second, you are invited to say, ‘that’s a stupid question.’ The point is that a rate is not necessarily a probability, and probability does not capture every kind of uncertainty.

In question 1, if the mechanism for deciding which of the 100 people will die in the next year is to select the tallest and shoot him or her, there is nothing random. There is then no *probability* that you will be the person who dies—you either are or are not the tallest person, just as you either are or are not named ‘Philip.’ If the mechanism for deciding who will die is to draw lots in a fair way and shoot the person who gets the short straw, that *is* reasonably modeled as random, and the probability that you are the person who dies is indeed 1%. The randomness is in the *method of selection*, not in the *rate*.

Rates and probabilities are not the same, and ignorance and randomness are not the same.

Cargo-Cult Confidence.²⁸

Suppose you have a collection of numbers, for example, a multi-model ensemble of climate predictions for global warming,²⁹ or a list of extinction rates for some cluster of previous studies. Take the mean and the standard deviation of this list of numbers. Report the mean as an estimate of something. Calculate the interval: mean, plus-or-minus 1.96 times the standard deviation. Claim that this is a 95% confidence interval or that there’s a 95% chance that this interval contains ‘the truth.’

That’s not a confidence interval for anything—and the probability statement is a further mangling of the interpretation of a confidence interval. If the collection of numbers were a random sample from some population and if that population had a Gaussian distribution (or if the sample size were large enough that you could invoke the central limit theorem), then the

²⁸ This is an allusion to *cargo cults* (https://en.wikipedia.org/wiki/Cargo_cult last accessed 16 November 2016), and, in particular, to Richard Feynman’s discussion of *cargo-cult science* (Feynman, R., 1974. *Cargo Cult Science*, *Engineering and Science*, 37(7), 10–13, <http://calteches.library.caltech.edu/3043/1/CargoCult.pdf> last visited 4 December 2017). The intervals are constructed using calculations that look like confidence interval calculations, but they are missing a crucial element: the data are not a random sample.

²⁹ This is essentially what IPCC does.

interval would be an approximate confidence interval for *something*. The probability statement would still be meaningless.

But if the list is not a random sample or a collection of measurements with random errors, there's nothing stochastic, period, and hence no confidence interval. A 95% confidence interval for a parameter is a number calculated from a random sample using a method that has at least a 95% probability producing an interval that includes the true value of the parameter. Once you've drawn the sample, everything is deterministic: even if you started with a random sample, the resulting interval either does or does not include the true value of the parameter.

Often, including in calculations I've seen in IPCC reports, people treat uncertainty in an estimate as if the truth is random and follows a probability distribution centered at the estimate. This is something like Fisher's fiducial inference,³⁰ which was abandoned by statisticians many years ago. The treatment is backwards: if the estimate is unbiased, the estimate has a probability distribution centered at the truth, not *vice versa*. Moreover, the truth is a fixed quantity, and once the estimate has been made, the estimate is also a fixed quantity. This practice is quantifauxcation.

Random versus haphazard.

Everyday language does not distinguish between 'random,' 'haphazard,' and 'unpredictable,' but the distinction is crucial for scientific inference. 'Random' is a very precise statistical term of art.

Here is an analogy: to know whether a soup is too salty, a very good approach is to stir the soup thoroughly, dip in a tablespoon, and taste the contents of the tablespoon. That amounts to tasting a random sample of soup. If I just dipped the spoon in without looking (and without stirring the soup), that would be a *haphazard* sample of soup. Those are very different processes. The second is *unpredictable* or *haphazard*, but it is not a random sample of soup, and it is not possible to quantify usefully the uncertainty in estimating the saltiness of the soup from a sample like that.

Notions such as probability, *P*-values, confidence intervals, etc., apply only if the data have a random component, for instance, if they are a random sample, if they result from random assignment of subjects to different treatment conditions, or if they have random measurement error. They do not apply to samples of convenience; they do not apply to haphazard samples; and they don't apply to populations. The mean and standard deviation of the results of a group of studies or models that is not a sample from anything does not yield *P*-values, confidence intervals, standard errors, etc. They are just numbers.

Freedman's Rabbit-Hat Theorem.

There are two rabbit axioms:

1. For the number of rabbits in a closed system to increase, the system must include at least two rabbits (preferably one male and one female),
2. There's no such thing as a negative rabbit.

From these, you can derive Freedman's Rabbit-Hat theorem:

³⁰ See, e.g., Seidenfeld, T., 1992. R.A. Fisher's Fiducial Argument and Bayes' Theorem, *Statistical Science*, 7, 358–368. <http://links.jstor.org/sici?sici=0883-4237%28199208%297%3A3%3C358%3ARAFFAA%3E2.0.CO%3B2-G> (last visited 16 November 2016)

***Theorem:** You cannot pull a rabbit from a hat unless at least one rabbit has previously been placed in the hat.*

***Corollary:** You cannot borrow a rabbit from an empty hat, even with a binding promise to return the rabbit later.*

Here are some applications. You can't turn a rate into a probability without assuming that the process was random in the first place: you can't conclude that a process is random without making assumptions that will amount to assuming that the process is random. Something has to put the 'randomness rabbit' into the hat. The existence of a rate doesn't do that. You have to get it from the 'physics' of the situation, or by assumption. The Corollary says you can't take a set of data, perform a test of randomness on the data and if they pass the test, conclude that the data are random. That amounts to borrowing a rabbit from an empty hat and claiming that it's okay because you returned the rabbit later. It doesn't work.

Let's discuss a few examples.

Example: Avian-Turbine Interactions.

A complaint that some have with wind power is that the turbines kill birds, in particular, raptors. This leads to a variety of questions. How many birds, and of what species? How big is the problem? What design and siting features of the turbines matter? Can you design turbines or wind farms in such a way as to reduce avian mortality?

I got slightly involved in this issue for the Altamont Pass wind farm in the San Francisco Bay area. To measure avian mortality from turbines, people walk around underneath the turbines and look for pieces of dead birds. The data aren't perfect. There is a background mortality rate of birds unrelated to wind turbines. Generally, you don't find whole birds, you find bird pieces. Is this two pieces of one bird or pieces of two birds? Carcasses decompose. Scavengers eat or take away pieces. And then there are problems of attribution: birds may be flung a distance from the turbine they hit, or they can be wounded and land some distance away. How do you figure out which turbine is the culprit?

Is it possible to make an unbiased estimate of the bird mortality from the turbines? Is it possible to relate that mortality to individual turbines and wind farms and to do it in a way that's reliable?

A standard stochastic model for the data is a 'zero-inflated Poisson process,' which is a mixture of a point mass at zero and a Poisson process. The extra mass at zero accounts for censoring: the data collection is likely to miss some of the deaths. There are many other models one might use instead. For instance, one might model the observations as the true count with errors that are dependent, not necessarily identically distributed, and not necessarily zero mean.

The management of the Altamont Pass wind farm hired a consultant, who modeled bird collisions with turbines as random, with a Poisson distribution with parameters that depend on properties of each turbine. The probability distribution of collisions is the same for all birds. Collisions are independent across birds, and the expected rates follow a hierarchical Bayesian model that relates the rate to variables including properties of the location and design of the turbine. Then he introduced additional smoothing to make the parameters identifiable. In the end, he came up with an estimator of the coefficients of the variables that the Poisson rates depend on.

If you try to describe in words what this model asserts about the world, things get weird: according to the model, when a bird approaches a turbine, in effect it tosses a biased coin. If the coin lands heads, the bird throws itself on the blades of the turbine. If the coin lands tails, the bird avoids the turbine. The chance the coin lands heads depends on the location and design

of the turbine, according to a pre-specified formula. For each turbine location and design, every bird uses a coin with the same chance of heads, and the birds all toss the coin independently.

Clearly, the person who invented this model chose to ignore that birds fly in flocks. Why are avian-turbine interactions random? I don't know. Why Poisson? I don't know. Why independent? I don't know. Why the same chance for all birds? I don't know. Why doesn't detecting a bird on the ground depend on how big the bird is, how tall the grass is, how long it's been since you last did a survey? I don't know.

Perhaps the most troubling thing is that consultant has changed the subject from 'how many birds does this turbine kill?' to 'what is the numerical value of some coefficients in this contrived Poisson?' This is no longer about birds. This is about a cartoon model—the model behind the curtain.

Earthquake Probability and Probabilistic Seismic Hazard Analysis.

Probabilistic seismic hazard analysis (PSHA) is the basis of seismic building codes in many countries. It's also used as a basis for deciding where it's safe to build nuclear power plants and nuclear waste disposal sites. PSHA seeks to estimate the probability of a given level of ground shaking (acceleration), for instance, a level that would damage the containment structure. It involves modelling earthquakes as occurring at random in space, time and with random magnitude. Then it models ground motion as being random, conditional on the occurrence of an earthquake of a given magnitude in a given place.

From this, PSHA claims to estimate 'exceedance probability,' the chance that the acceleration in some particular place exceeds some tolerable level in some number of years. In the US, building codes generally require structures to withstand accelerations that will occur with probability of 2% or greater in 50 years. PSHA came out of probabilistic risk assessment, which originated in aerospace and nuclear power, primarily.

A big difference between PSHA and these other applications is even if you've never launched a vehicle to the moon before, the spacecraft is an engineered system and you have a pretty good idea of its properties, the environment it's operating in, and so forth and so on. Even if you've never built a nuclear reactor before, you know something about concrete. You know something about steel. You know something about nuclear physics. You know something about pressure vessels.

We know very little about earthquakes, frankly, other than their phenomenology. We don't really understand the physical generating processes. We don't know in detail how they occur. There's a big difference between an engineered system and a natural system that is inaccessible to us.

PSHA models earthquakes as a marked stochastic process with known parameters. The fundamental relationship in PSHA is that the probability of a given level of ground movement in a given place is the integral over space and magnitude of the conditional probability of that level of movement given that there's an event of a particular magnitude in a particular place times the probability that there's an event of a particular magnitude.

This is just the law of total probability and the multiplication rule for conditional probabilities, but where's it coming from? That earthquakes occur at random is an assumption, not a matter of physics. Seismicity is complicated and unpredictable. But that's *haphazard*, not *random*. The standard argument to calibrate the PSHA fundamental relationship requires conflating rates with probabilities. For instance, suppose a magnitude eight event has been observed to occur about once a century in a given region. PSHA would assume that, therefore, the chance of a magnitude 8 event is 1% per year.

That's wrong, for a variety of reasons. First, there's an epistemic leap from a rate to the existence of an underlying random process that generated the rate, as discussed above. Second, there's an assumption that the process is uniform, which contradicts the observed clustering of seismicity in space and time. Third, this ignores the fact that the data at best give an estimate of a probability (if there is such a probability), not the exact value.

PSHA relies the metaphor that earthquakes occur as if in a casino game. According to the metaphor, it's as if there is a special deck of cards, the earthquake deck. The game involves dealing one card per time period. If the card is blank, there is no earthquake. If the card is an eight, there is a magnitude eight earthquake. If the card is a six, there's a magnitude 6 earthquake, and so forth.

There are tens of thousands of journal pages that, in effect, argue about how many of each kind of card there are in the deck, how well shuffled the deck is, whether after each draw you replace the card in the deck and shuffle again before dealing the next card, whether you add high-numbered cards if you haven't drawn any in a while, and so on. The literature, and the amount of money spent on this kind of work, are enormous—especially given that it has been unsuccessful scientifically. Three of the most recent destructive earthquakes internationally were in regions that seismic hazard maps said were relatively safe.³¹ This should not be surprising, because the enterprise is based on a metaphor, not on physics.

Here's a different metaphor: earthquakes occur like terrorist bombings. You don't know when or where they're going to happen. You know that they're going to be large enough to hurt people when they do happen, but not how large. You know some places are likely targets. But there's no probabilities. You might choose to *invent* a probability model to see whether you can do better law enforcement or prevention, but that's different from saying that a terrorist is rolling a die to decide when and where to strike and with how large a bomb. That is a very different thing. It's not a generative model. It's attempting to be a predictive model.

What advantages might there be to using a casino metaphor for earthquakes? It would be an apt metaphor if the physics of earthquakes were stochastic—but it isn't. It would be apt if stochastic models provided a compact, accurate representation of earthquake phenomenology, but they don't: the data show that the models are no good.³² The metaphor might be apt if the models led to useful predictions of future seismicity, but they don't.³³

You can predict earthquakes quite well with a rule of the form, 'If there is an earthquake of a magnitude X or greater, predict that there is going to be another one within Y kilometers within Z days.'³⁴ Such rules perform quite well, without relying on invented stochastic mumbo jumbo.

³¹ Stein, S., Geller, R.J., and Liu, M., 2012. Why earthquake hazard maps often fail and what to do about it, *Tectonophysics*, 562–562, 1–25.

³² See Luen, B., and Stark, P.B., Poisson tests of declustered catalogs, *Geophysical Journal International*, 189, 691–700; Luen, B., 2010. *Earthquake prediction: Simple methods for complex phenomena*, Ph.D. Dissertation, Department of Statistics, University of California, Berkeley, UMI Number 3449030.

³³ Luen, B., 2010. *Earthquake prediction: Simple methods for complex phenomena*, Ph.D. Dissertation, Department of Statistics, University of California, Berkeley, UMI Number 3449030.

³⁴ Luen, B. and P.B. Stark, 2008. Testing Earthquake Predictions. IMS Lecture Notes—Monograph Series. Probability and Statistics: Essays in Honor of David A. Freedman, 302–315. Institute for Mathematical Statistics Press, Beachwood, OH. Reprint: <http://arxiv.org/abs/0805.3032>; Luen, B., 2010. *Earthquake prediction: Simple methods for*

Climate Models.

The IPCC wants to treat uncertainties as random:

...quantified measures of uncertainty in a finding expressed probabilistically (based on statistical analysis of observations or model results or expert judgment).

... Depending on the nature of the evidence evaluated, teams have the option to quantify the uncertainty in the finding probabilistically. In most cases, level of confidence...

... Because risk is a function of probability and consequence, information on the tails of the distribution of outcomes can be especially important... Author teams are therefore encouraged to provide information on the tails of distributions of key variables...³⁵

In these few sentences, the IPCC is asking or requiring people to do things that don't make sense and that don't work. As discussed above, subjective probability assessments (even by experts) are generally nonsense, and subjective confidence is unrelated to accuracy. Mixing measurement errors with subjective probabilities doesn't work. And climate variables have unknown values, not probability distributions.

Cargo-cult confidence confusion seems to be common in IPCC reports. For instance, the 'multi-model ensemble approach' the IPCC uses involves taking a group of models, computing the mean and the standard deviation of their predictions, then treating the mean as if it is the expected value of the outcome (which it isn't) and the standard deviation as if it is the standard error of the natural process that's generating climate (which it isn't).³⁶

This is wrong all the way down.

Simulation.

In some fields—physics, geophysics, and climate science in particular—there is a popular impression that probabilities can be estimated in a 'neutral' way by doing Monte Carlo simulations: just let the computer generate the distribution. Monte Carlo simulation is a way to

complex phenomena, Ph.D. Dissertation, Department of Statistics, University of California, Berkeley, UMI Number 3449030.

³⁵ Mastrandrea, M.D., C.B. Field, T.F. Stocker, O. Edenhofer, K.L. Ebi, D.J. Frame, H. Held, E. Kriegler, K.J. Mach, P.R. Matschoss, G.-K. Plattner, G.W. Yohe, and F.W. Zwiers, 2010. *Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties*. Intergovernmental Panel on Climate Change (IPCC). <https://www.ipcc.ch/pdf/supporting-material/uncertainty-guidance-note.pdf> (at p.2, last accessed 16 November 2016) The authors of the paper have expertise in biology, climatology, physics, economics, ecology, and epidemiology. To my surprise—given how garbled the statistics is—one author, Francis Zwiers, is a statistician.

³⁶ The IPCC also talks about simulation errors being (to some extent) 'independent,' which presupposes that such errors are random. https://www.ipcc.ch/publications_and_data/ar4/wg1/en/ch10s10-5-4-1.html (last accessed 16 November 2016) But modeling errors are not random—they are a textbook example of systematic error. And even if they were random and independent, averaging would tend to reduce the variance of the result, but not necessarily improve accuracy, since accuracy depends on bias as well.

substitute computing for hand calculation. It is not a way to discover the probability distribution of anything. It's a substitute for doing an integral, not a way to uncover laws of nature.

Monte Carlo doesn't tell you anything that wasn't already baked into the simulation in the first place. The distribution of the output comes from assumptions in the input (modulo bugs). It comes from what you program the computer to do. Monte Carlo reveals the consequences of your assumptions, not anything new. The randomness is an assumption. The rabbit goes into the hat when you build the model and write the software. The rabbit does not come out of the hat without having gone into the hat first.

The Rhodium Group Climate Prospectus.

The Bloomberg Philanthropies, the Office of Hank Paulson, the Rockefeller Family Fund, the Skoll Global Threats Fund, and the TomKat Charitable Trust funded a study³⁷ that purports to predict various impacts of climate change.

The report starts somewhat circumspect:

While our understanding of climate change has improved dramatically in recent years, predicting the severity and timing of future impacts is a challenge. Uncertainty around the level of greenhouse gas emissions going forward and the sensitivity of the climate system to those emissions makes it difficult to know exactly how much warming will occur and when. Tipping points, beyond which abrupt and irreversible changes to the climate occur, could exist. Due to the complexity of the Earth's climate system, we don't know exactly how changes in global average temperatures will manifest at a regional level. There is considerable uncertainty...

But then,

In this climate prospectus, we aim to provide decision-makers in business and government with the facts about the economic risks and opportunities climate change poses in the United States.

Yep, the 'facts.' They proceed to estimate the effect that climate change will have on mortality, crop yields, energy use, the labor force, and crime, *at the level of individual counties in the United States through the year 2099*. They claim to be using an 'evidence-based approach.' Their approach is 'evidence-based' in the same sense alien abduction movies are based on a true story.

Among other things, the prospectus predicts that violent crime will increase just about everywhere, with different increases in different counties. How do they know? Generally, on hot days, there's more crime than on cool days. Fit a regression model to the increase. Assume that the fitted regression model is a response schedule. Input the average temperature change predicted by the climate model in each county; out comes the increase in crime rate.

Think about this for a heartbeat. Even if you knew exactly what the temperature and humidity would be in every cubic centimeter of the atmosphere every millisecond of every day, you would have no idea how that would affect the crime rate in the U.S. next year, much less in

³⁷ Houser, T., Hsiang, S., Kopp, R., and Larsen, K., 2015. *Economic Risks of Climate Change: An American Prospectus*, Columbia University Press, NY. See also http://riskybusiness.org/site/assets/uploads/2015/09/RiskyBusiness_Report_WEB_09_08_14.pdf (last accessed 16 November 2016)

2099, much less at the level of individual counties. And that's before you factor in the uncertainty in climate models, which is astronomical, even for globally averaged quantities, much less variables at the level of individual counties.³⁸ And it's also before you consider that society is not a constant: changes in technology, wealth, and culture over the next 100 years surely matter.

Global circulation models (climate models) are theorists' tools, not policy tools: they might help us understand climate processes, but they are not a suitable basis for planning, economic decisions, and so on.³⁹ The fact that intelligent, wealthy people spent a lot of money to conduct this study and that the study received high-profile coverage in *The New York Times* and other visible periodicals speaks to the effectiveness of quantifauxcation as a rhetorical device. I can't help but wonder whether there is a hidden agenda, since the predictions are obviously garbage.

Preproducibility.

Preproducibility is another neologism. Why do we need it? There are many related terms with overloaded and contradictory meanings in different fields, such as replicable, reproducible, repeatable, confirmable, stable, generalizable, reviewable, auditable, verifiable, and validatable.

Scientific conclusions usually assume that some *ceteris* are *paribus*. Andrea Saltelli says *ceteris* are never *paribus*, and I agree. Would you get a similar result if you repeated the experiment in the same lab? If somebody else repeated it elsewhere? If it were done in 'similar circumstances?' Would you get the same graphs and numbers if somebody repeated your data analysis and so forth?

With respect to what changes is the result stable? Changes of what size, and how stable? What *ceteris* need to be *paribus* for a result to hold? Karl Popper said, 'Science may be described as the art of systematic over-simplification – the art of discerning what we may with advantage omit'.

I claim that the desired level of abstraction or generalization—the *ceteris* that need not be *paribus*—defines the scientific discipline you're working in. If you want to generalize to all time and all universes, you're doing mathematics. If you want to generalize to our universe, you're doing physics. If you want to generalize to all life on Earth, you're doing molecular and cell biology. If you want to generalize to all mice, you're doing murine biology. If you want to generalize to C57BL/6 mice, I don't know what you're doing. If you only care about one mouse in one lab in one experiment on that day, I'm not convinced that you're doing science at all.

There is a wonderful essay by J.B.S. Haldane,⁴⁰ in which he writes:

You can drop a mouse down a 1,000-yard mineshaft and, on arriving at the bottom, it gets a slight shock and walks away, provided that the ground is fairly soft. A rat is killed; a man is broken; a horse splashes. For the resistance presented to movement by the air is proportional to the surface of the moving object.

³⁸ See, e.g., Regier, J.C. and P.B. Stark, 2015. Uncertainty Quantification for Emulators. *SIAM/ASA Journal on Uncertainty Quantification*, 3, 686–708. doi:10.1137/130917909, Reprint: <http://epubs.siam.org/toc/sjuq a3/3/1>

³⁹ See, e.g., Saltelli, A., P.B. Stark, W. Becker, and P. Stano, 2015. Climate Models as Economic Guides: Scientific Challenge or Quixotic Quest?, *Issues in Science and Technology*, Spring 2015. Reprint:

<http://www.stat.berkeley.edu/~stark/Preprints/saltelliEtal15.pdf>

⁴⁰ Haldane, J.B.S., 1926. On being the right size, *Harper's Magazine*.

Is this physics? Is this biology?⁴¹

The tolerable variation in experimental conditions depends on the inference that you'd like to make. If variations in conditions that are irrelevant to the discipline change the result, you have a replicability problem: the outcome doesn't have the right level of abstraction. It's probably too general.

There are two joined issues relating to replicability of scientific results.

1. Can the experiment be repeated?
2. When the experiment is repeated, are the results the same?

If an attempt to replicate does not give the same results, *why* did it fail? It could be that the effect is intrinsically variable or intermittent, that the result is a statistical fluke or 'false discovery,' or that something that mattered was different, so the claim needs to be qualified. A result is *preproducible* if it has been described down to the level of things 'that you may with advantage omit.' If you haven't adequately described what you did, including all the data cleaning and data processing, you are simply asking others to trust you, not providing a basis for others to evaluate your claims.

Science should not require trusting authority: it should be *show me*, not *trust me*.⁴² The Royal Society, and Boyle's conception of Science, originally required the evidence to be public⁴³—a demonstration in front of an educated audience—for a scientific claim to be considered established. From that evolved a style of reporting experiments that helped the reader imagine being in the room in which the demonstration took place and observing the demonstration. Details mattered: the idea was to be able to recreate the experiment, at least in one's mind. That standard has eroded, in part (perhaps) because the apparatus for collecting and analyzing data has become increasingly complex, making it hard to describe in sufficient detail what was done. Instead of providing evidence, many scientific publications have devolved into advertising.⁴⁴ Publications simply say, 'I discovered X.' Far too often, publications don't even say what the author claims to have done to discover X.

Here is a list of questions that would be nice to have answers to when evaluating any scientific claim:

- Were the materials (organisms), instruments, procedures, and conditions specified adequately to allow repeating the data collection?
- What are the raw data?
- How were they collected or selected?
- How were the raw data processed to get the 'data' that the claims are based on?

⁴¹ Lord Rutherford said, 'All science is either physics or stamp collecting.'

⁴² Stark, P.B., 2015. Science is "show me," not "trust me." Berkeley Initiative for Transparency in the Social Sciences, <http://www.bitss.org/2015/12/31/science-is-show-me-not-trust-me/> (last accessed 16 November 2016)

⁴³ Well, sort of public: visible to members of the Society and the occasional royal. See, e.g., Shapin, S. and Schaffer, S., 1985. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*, Princeton University Press, Princeton, NJ

⁴⁴ See Donoho, D.L., 2010. An invitation to reproducible computational research, *Biostatistics*, 11, 385–388, and references therein.

- How were the processed data analyzed?
- Was the data analysis described adequately to allow others to check and repeat it?
- Are code and data available to re-generate figures and tables?
- Is the code readable and checkable?
- Is the software build environment specified adequately?
- Was the analysis performed the right analysis to perform?
- Was it performed correctly?
- Were there ad hoc aspects to the analysis?
- What would have happened if different choices had been made?
- What other analyses were tried? What happened?
- How was multiplicity treated?
- Can others use the procedures and tools?
- Were the results reported correctly?
- How generally do the results hold? how stable are the results to perturbations of the experiment?
- Most of all: *What is the evidence that the result is correct?*

Software matters.

According to a practitioner:

Rampant software errors undermine scientific results. Errors in scientific results due to software bugs are not limited to a few high-profile cases that lead to retractions and are widely reported. Here I estimate that in fact most scientific results are probably wrong if data have passed through a computer, and that these errors may remain largely undetected. The opportunities for both subtle and profound errors in software and data management are boundless, yet they remain surprisingly unappreciated.⁴⁵

How can we do better? The first step is to adopt tools used in the software development world: revision control systems, standard documentation, coding standards and conventions, pair programming, issue tracking, code reviews, unit testing, code coverage testing, integration testing, regression testing, scripted analyses.⁴⁶ If you're using Excel, go back to square one and start over. I don't care what you *claim* you did: your analysis is not trustworthy.⁴⁷ If you're using a spreadsheet for anything more than data entry, your answer is probably wrong. Spreadsheets make it very easy to make mistakes, and very hard to find them, because they conflate input, programming, output, and presentation. They make debugging difficult. They make unit testing difficult. They make replication difficult. They make code review difficult.

Everyone should see the European Spreadsheet Risk Interest Group website (www.eusprig.org), which includes a compendium of disaster stories resulting from

⁴⁵ Soergel, D.A.W., 2014. <http://f1000research.com/articles/3-303/v2> (last accessed 16 November 2016)

⁴⁶ See Wilson, G., D.A. Aruliah, C.T. Brown, N.P. Chue Hong, M. Davis, R.T. Guy, S.H.D. Haddock, K.D. Huff, I.M. Mitchell, M.D. Plumbley, B. Waugh, E.P. White, and P. Wilson, 2013. Best Practices for Scientific Computing. <http://arxiv.org/pdf/1210.0530v4.pdf> (last accessed 16 November 2016)

⁴⁷ For a checklist of practices that promote reproducibility, see Stark, P.B., 2015. Science is “show me,” not “trust me.” Berkeley Initiative for Transparency in the Social Sciences, <http://www.bitss.org/2015/12/31/science-is-show-me-not-trust-me/>

spreadsheet errors. Their top examples include the Reinhart and Rogoff study, that JP Morgan understated their value-at-risk calculations for Basel II compliance, that the Olympic International Organizing Commission oversold a skating event by 10,000 tickets, that a county in Tennessee and a city in Wisconsin made errors costing them millions, and many more. According to KPMG and PwC, two major consulting firms, over 90% of corporate spreadsheets have errors.⁴⁸

Spreadsheets are problematic in general, but Microsoft Excel in particular has been plagued by computational bugs, including bugs in addition, multiplication, random number generation, and in statistical routines. Some of the bugs persisted for several versions of the program.⁴⁹

The bug in the pseudorandom number generator (PRNG) survived for several generations of Excel (supposedly it was fixed in 2010, but it's not possible to check). Microsoft claimed that Excel implemented the Wichmann-Hill PRNG. It's not a very good PRNG (among other things, it has a short period, inadequate for serious statistics), but it *is* supposed to give numbers between zero and one. Some versions of Excel occasionally gave negative numbers, so they had a bug in the half a dozen or so lines of code it takes to implement the Wichmann-Hill PRNG.

But Excel does not allow you to set the seed for the PRNG, so you can't replicate the problem. And since you can't set the seed, you can't replicate anybody else's Excel simulations or samples. Stress tests of the international banking system are largely Excel simulations, done on very large computers. Be worried.

Relying on spreadsheets for important calculations is like driving drunk. No matter how carefully you do it, a wreck is likely.

Suggestions and recommendations.

What should we do (differently)? Whether your role is consumer or producer, I recommend:

- Think about where the data come from and how they happened to become the sample. Are they a random sample? A sample of convenience? The result of a randomized, controlled experiment? An observational study? Are there controls? Was there any intervention? What's the response rate? Was there self-selection? Did the sampling frame match the population of interest?

⁴⁸ http://www.theregister.co.uk/2005/04/22/managing_spreadsheet_fraud

⁴⁹ See, e.g., McCullough, B.D., and B. Wilson, 2005. On the accuracy of statistical procedures in Microsoft Excel 2003, *Computational Statistics & Data Analysis*, 49, 1244–1252, <http://www.sciencedirect.com/science/article/pii/S0167947304002026>; Knüsel, L., 2005. On the accuracy of statistical distributions in Microsoft Excel 2003, *Computational Statistics & Data Analysis*, 48, 445–449,

<http://www.sciencedirect.com/science/article/pii/S0167947304000337>; McCullough, B.D., and Heiser, D.A., 2008. On the accuracy of statistical procedures in Microsoft Excel 2007, *Computational Statistics & Data Analysis*, 52, 4570–4578,

<http://www.sciencedirect.com/science/article/pii/S0167947308001606>;

McCullough, B.D., 2008. Microsoft Excel's 'Not the Wichmann-Hill' random number generators, *Computational Statistics & Data Analysis*, 52, 4587–4593.

also <http://www.cnet.com/news/beware-of-a-bug-in-excel-when-doing-addition/> and http://www.theregister.co.uk/2007/09/26/excel_2007_bug/ (both last accessed 16 November 2016)

- Are the data available to others? Do they pass a “sniff test” for silliness? If the data are secret, be skeptical of the results: only the authors can check the accuracy of their computations.
- Assumptions matter: Enumerate the assumptions of the experiment and data analysis. Check those you can; flag those you can't. Which are plausible? Which are plainly false? How much might it matter?
- If there is a probability statement, consider how probability came into the problem. Is anything actually random? Was there a random sample? A randomized assignment to treatment or control? Random measurement error? Is the probability in any sense “calibrated” against empirical data? Is the statement testable? Is the probability a subjective measure of certainty, or is it tied more closely to data or the “physics” of the problem? Is the probability coming from an invented model? Is the probability just metaphorical? In what sense is the phenomenon like a casino game?
- Be skeptical of hypothesis tests, standard errors, confidence intervals, and the like in situations that do not involve real randomness from random sampling, random assignment, or random measurement error. The numbers almost certainly result from “cargo-cult” calculations and have little to do with how well the data support the conclusions.
- Errors never have a normal distribution. The consequence of pretending that they do depends on the situation, the science, and the goal. Think about the sensitivity of the results to distributional assumptions. Can the distributional assumptions be relaxed or avoided by using a different analysis?
- Pay attention to sources of systematic error. Do the data measure what they claim to measure? Has there been a change of subject? Are the data what was actually measured, or do they result from processing some underlying observations? If the latter, what are the assumptions behind that processing? Are they reasonable? What assurance is there that the processing was performed correctly?
- Be concerned about software bugs. Test your own code (and publish your code, including your tests of that code). Investigate how others’ code was tested. Script your analyses: don’t rely on point-and-click tools. Use unit tests, integration tests, and regression tests. Check the coverage of your tests. Consider pair programming. Remember that there’s always a bug, even after you find the last bug.

Checklist.

Here is a brief diagnostic checklist I find helps me improve the reproducibility of my data analysis:

- Did you use a spreadsheet for computations? (Don’t!)
- Did you script your data analysis, including data cleaning and munging? Did you make the script available?
- Did you document your code so that others can read and understand it?
- Did you record and report the versions of the software you used, including library dependencies?
- Did you write tests for your code?
- Did you check the code coverage of your tests?
- Did you use open-source software for your analysis? If not, is there an open-source equivalent to the software you used?
- Did you report all the analyses you tried (transformations, tests, selections of

variables, models, etc.) before arriving at the one you chose to emphasize?

- Did you make your code (including tests) available?
- Did you make your data available?
- Did you record and report the data format?
- Is there an open-source tool for reading data in that format?
- Did you provide an adequate data dictionary?
- Did you provide free public access to your work, e.g., by publishing in an open-access journal?

Pledge.

If you want to push science back in the direction of reproducibility, here are some pledges that might shift the tide if enough people followed them:

1. *I will not referee (or accept for publication) any article that does not contain enough information to check whether it is correct.*⁵⁰
2. *I will not submit for publication any article that does not contain enough information to check whether it is correct.*
3. *I will not publish in any journal that does not allow me to publish enough information to check whether my work is correct.*
4. *I will not serve as an editor or referee for any journal that does not allow authors to provide enough information to check whether their work is correct.*
5. *I will not rely on any article published after [date] that does not contain enough information to check whether it is correct.*

Acknowledgment.

I am grateful to Robert Geller, Jim Rossi, and Andrea Saltelli for helpful suggestions.

⁵⁰ The 2015 *Peer Reviewer's Openness Initiative* expands on this: <https://opennessinitiative.org/the-initiative/> (last accessed 16 November 2016)