A philosophical standpoint for the practice of statistical inference

Robert L Grant
(writing in a personal capacity)

c/o BayesCamp Ltd
16 City Business Centre, Hyde Street,
Winchester, SO23 7TA, United Kingdom

May 16, 2022

Version 1.3 — `git commit 24683b7`

Abstract

There are different sets of philosophical underpinnings — definitions and assumptions — in use for statistical inference and associated activities such as communication and teaching. They have implications for which analytical methods can be used, and how they can be communicated. A statistician therefore should develop a personal set of defensible philosophical principles. This helps their audience to avoid confusion but also empowers them to justify and defend their decisions. More widespread work to establish such deep and holistic standpoints could help the profession, and data scientists, to move beyond a tribalism that prevents impactful, collaborative, interdisciplinary work.

In this informal and occasionally tongue-in-cheek document, I explore the philosophical background to various practical problems in statistics that cause me trouble, and attempt to bring the aspects together into one personal standpoint. This leads to a "justified eclecticism" with a novel division of statistical methods into four quadrants, based on estimands and the order of identification and evaluation of parameter values.

This depends on a surprisingly wide range of foundations. First, I refute some silly ideas about frequentism that are often found in the wild, and get down to the topic of the long run frequency. How long is long? I explain why I think the sensible answer is "it depends", and how to go about justifying this choice, alongside the population for inference.

I explore Lyotard's notion of performative science, and adopt a relevant definition of modernism and postmodernism from Augé. I start

to discuss communication and priors, before making a case for realistically complex models, constrained (rendered identifiable) by exogenous theory (experience, previous evidence, qualitative data ...).

Now, competing definitions of probability are shown to boil down to the metaphysical question of realism or nominalism of the reference class that defines the denominator for the long run frequency. Personally, I reach probabilistic nominalism via two arguments: a weaker one via resemblance nominalism and a stronger one via existence monism, but there are other routes.

I use Yablo's aboutness to assert that statements using aleatory-only and aleatory-epistemic conceptions of probability share meaning if the estimand is the data generating process (DGP), while only the aleatory-epistemic serves for estimating belief in the DGP. This subdivides statistical inference at a different place to the common frequentist-versus-Bayesian concept.

Across this, we can split methods according to whether parameter values are identified then evaluated (as in Neyman-Pearson-Wald testing or Bayes factors), or evaluated then identified (as in maximum likelihood estimation, Fisher testing, or Markov chain Monte Carlo).

Emphasis is given to serving the audience, through communication, compromise and realistically complex models, rendered identifiable through constraints from exogenous theory. Complexity, Lipton's inference to the best explanation and Odell's "doing nothing" help to complete a guide to what to do and how to do it. The standpoint is necessarily a snapshot in time and may evolve; I close with my to-do list for advancing it.

Contents

1 Introduction

> *"Oh. You've gone wacky."*
> — A fellow data visualisation enthusiast, in conversation with the author at London Stata Users' Group 2013.

Statisticians sometimes argue among themselves about the right way to turn data into insights. There are methodological debates, and also philosophical ones, and among the latter, some have grumbled on for a very long time. This paper is about those fundamental questions of what one is trying to do, and how one might do it. It is opinionated, because there are no global answers to these problems.

In writing it, I have assumed some familiarity with the basics of statistics, and perhaps also with the everyday problems of negotiating an agreed set of aims, methods, findings and conclusions among collaborators who do not exactly share one's level of understanding or standpoint, not to mention communicating findings to an audience with low statistical literacy and confirmation bias.

I try to explain the statistical concepts which are not covered in a basic course (usually, these take the form of Snedecor's august curriculum[*1]), as well as all of the relevant philosophy, and to signpost to further reading where it might be helpful. Nevertheless, a statistician who has not studied or read a lot of philosophy will probably have to look up a lot of the ideas I pass through. It took me about 13 years to get from there to here, but I still keep a[†] copy of Godfrey-Smith's excellent overview of philosophy of science.[2] The material I have had to cover is unfortunately wide-ranging, and gets difficult at points, but to paraphrase Ken Hom, if I can do it then so can you.

I write about "statisticians", but lots of people analyse data, and I recommend this sort of contemplation to them all.

Given that one may be challenged on one's approach to analysis at any moment, it seems incumbent upon the statistician who attains an intermediate level of skill with his or her subject to reflect on these fundamental issues, and to seek a clear set of personal choices and justifications. Then, others may still disagree, but the statistician will not be unduly influenced or thrown into doubt.

"Missing data cannot have a probability distribution or confidence interval about its values, because it is meaningless to say it has a probability of lying in some range. It either is there or is not. There is no probability, only ignorance on your part." Thus spake a lecturer when, as a student, I expressed curiosity about a method for accommodating missing data called

[*] Chapters 1–7, 10 and 11, contents of which can be read online at https://archive.org/details/in.ernet.dli.2015.5515/page/n11/mode/2up

[†] that's how I'm livin'

multiple imputation.[3] I did not know the history at the time, but they were drawing on an old debate about the ways in which one may use probability. As we learn our craft, most statisticians will have such encounters. At first, we are pulled this way and that, absorbing the view we just heard expressed so forcefully. We try to reconcile everything we have heard from experienced and wise statisticians, and some confusion results, as we find that their various pronouncements are not entirely compatible with one another.

Introductory courses do not admit to these fundamental disagreements, which is the reason for the confusion (it's not your fault). Later, the statistician learns to recognise the various tribes and their dogmas, and finally, to consider the arguments in depth and arrive at their own set of principles. The lecturer who dismissed multiple imputation as the work of "mad Bayesians" had glossed over one of those fundamental disagreements: their statement is true *if* you assume that probability only represents long-run frequencies.[‡]

But why assume that? It is a philosophical point, one that you are free to agree or disagree with. It cannot be proven or disproven, or shown to be superior or inferior to an alternative interpretation by some kind of methodological study. And yet, we build our statistical practices, with all their mathematical rigour, on top of it.

Probability is a little strange in this respect, because usually mathematics begins with definitions and proceeds to theorems that can be applied to problems. In this case, probability is something intuitive, which is already used to tackle problems, but not very well. Time and time again, we have learnt how humans do not naturally think clearly in probabilistic terms. Seeking to help this, mathematicians worked backwards to make a definition — and they have never agreed on that.

Our goal should be to decide whether we agree or not with these philosophical points, and then, best of all, to know *why* we hold our standpoint, in such a way that all our views are justifiable and coherent with each other.

The rest of this paper visits some of these problems and sets out my own approach. It is a record of why I do what I do, hopefully a stimulus to others to do the same, and perhaps a demonstration of the principle that to practice statistical inference *is* to practice philosophy of science, so if we write a Methods section, we had better have a Philosophy section stored away somewhere. This paper is an attempt at writing my own Philosophy section.

[‡]It might be that my lecturer had never thought about it that much. Most statisticians do not. We are human, after all, and place great store in *ipse dixit*.

2 Background

2.1 Inductive and deductive statistics

An *inductive* inference is from specific observations to general principles.[2] For example, upon finding the cheese gone in the morning, you conclude your house has mice. Induction is tricky, though: in order to weigh up one inference over another (your kids crept downstairs and ate it), we need more information from outside the original observation; we'll come back to that.

You might think that this sounds like statistics, and that is broadly right: we are in the business of turning observations into generic insights about how the world works. Statistics mostly is about formalising induction so that it becomes more reliable. The form applied is that of the probability distribution, a mathematical tool which we tackle in more depth in Section 2.2. By applying an assumed probability distribution, we get rigorous inductive inferences backed by mathematical theorems. The cost of this is that we have to reduce our inferences to simplified quantitative summaries rather than rich, multifaceted stories, and of course they will only be rigorous if our assumptions hold.

Deduction, in contrast, is an attempt to learn about the general principles by predicting specific observations from them, then looking to see if those observations actually occur. This is the basis of a lot of physical sciences going back to the Enlightenment, or at a pinch, to Aristotle: lab-book stuff. It brings in two concepts that were actually quite foreign to early statistics: controlled experiments and falsification.

If you can control the circumstances of data collection very carefully, then you can eliminate all causes but the one you are interested in. This way, you make the experiment harder and the analysis easier. Running a t-test on a randomised, controlled, double-blind clinical trial is much simpler than fitting a multilevel model with imputation of missing data to a retrospective dataset. But sometimes you don't have a choice. And who wants to do easy stats all the time?

History tells a story of statistical fashion switching between induction and deduction, and not necessarily ending up making much sense when viewed in its entirety. First, there was probability theory applied to strictly controlled repeated events (you might almost call them experiments), mainly gambling (the Bernoullis, Laplace). Then, there was correlation between pairs of variables (Pearson the elder), which is evidently different because it is inductive.

Then, there was some rebellion to use experiments rather than slap any handy data together and draw strong conclusions (Fisher). The new output from this was the p-value, which uses a deductive approach but keeps the final decision a little bit subjective and soft round the edges. Fisher suggested $p < 0.05$ as a threshold for evidence, but not for proof. He suggested that

multiple experiments, all returning $p < 0.05$, could constitute something close to proof. He also left the door open to move the p threshold to a different level, depending on circumstances.

Next, there was a move by Neyman and Pearson (the younger), and later Wald, to really firm up the deductive aspects. In this framework, you must specify your hypothesis up front, not cherry-pick, and you must declare the results significant or not at a pre-specified threshold. On this basis, they were able to put deductive statistics in a decision-theoretic framework, and hence we have Type 1 and Type 2 errors, power and so forth.

This is good practice in deduction for binary decisions (the doctor can only prescribe one of the two competing drugs, the general either authorises the cruise missile or doesn't), but leaves something lacking if the goal is more exploratory. Everyone who has encountered statistics for any length of time knows these problems: statistical significance says little about practical significance, and knowing that the null hypothesis is false does not get you any further forward with saying what the most likely population value is, let alone what the chance is of it lying in some range.

More flexible inferences were done in the decades that followed by working with likelihood: the probability model for the data-generating process (DGP), turned on its head so that it tells you how well different putative parameter values (plus assumed DGP) fit the data. Broadly speaking, there is a trade-off between the rigour of Neyman-Pearson and the flexibility to craft complex models for how the data came into being. We'll expand on this in Section 6.

2.2 Probability

The lecturer I mentioned above subscribed to frequentism, which defines probability only as the proportion of events from a long sequence, identically repeated, which match some particular conditions. This leads to random sampling, sampling distributions, standard errors, the central limit theorem, all kinds of useful stuff.[4] It solves some kinds of problems very well.

It has a rock-solid philosophical foundation because that connects directly with the mathematical foundation. Such a long sequence of n events will indeed return a proportion p, from which we can make an estimate of the "true" proportion in the population whence the sample was drawn: $\hat{\pi}$ (not every sample statistic relates as directly to the estimate as this, but many are simple plug-ins), and that will tend to a certain limit (the "true" population parameter π, which we see but through a glass darkly) as n tends to infinity. π is also immediately defined as the probability of the next event matching the conditions, so it functions not only as a description of the population whence the data were drawn, but also has predictive power.

Frequentists typically stop there. Other people have devised other meanings of probability[5] but only frequentism connects objectively to the maths.

Whether you like the idea of probability as a kind of ineffable propensity for data to come out a particular way, or as a degree of personal belief, or as a calculus of fuzzy logic, you have to admit two things. Firstly, the definition makes no odds (no pun intended) to the mathematics of probability. Once you adopt that definition, everything else that follows in practically using probability is the same, and the same mathematical theorems hold. Their outputs might represent different things, but the maths still works just fine. Secondly, all other definitions extend beyond the long-[eternal-]run of identically repeated events, and so allow for uncertainty arising from sources other than sampling error.* I encounter from time to time some folksy interpretations of frequentism, which I will debunk in footnotes hither and yon. This is not to say that I reject frequentism.

The sampling error is called *aleatory* uncertainty (from the Latin for dice, like the American expression *it's just a crap shoot*) and anything else, like the uncertainty around missing data, is *epistemic* uncertainty (from the Greek for knowledge, like the American expression *I don't know s*** from Shinola*). Mainly, I am concerned with *Bayesian* statistics, which uses probability to represent both forms of uncertainty, and returns probability distributions of the unknowns, given the data, $P(\theta|X)$. We call these *posterior* distributions because they represent our information (or belief, or whatever version of probability you prefer) about θ *after* we have encountered the data. Hence the idea of updating a prior distribution $P(\theta)$ by multiplication with the likelihood $L(\theta|X) = P(X|\theta)$ to get the posterior (*modulo* a normalising constant).

We know that there are parameters — unknown values sought by our analyses — which, if we were to repeat our data collection process n times ($n \in \mathbb{N}$), would always take different values, drawn at random from their aleatory sampling distributions. Everybody calls these parameters.

Also, there are unknown values which never take new values from an aleatory sampling distribution, even for $n = 1$, such as missing data, so it is entirely consistent with the premiss of frequentism that inference on these

*Some practitioners of frequentist statistics appeal to the question of whether an unknown (like a missing datum) has a value known to someone else, but this opens the door to a paradox. A measurement recorded by a machine, and then lost by the researcher, could be imputed using probability, because nobody ever knew it. But, as soon as someone finds the lost file, the analysis and all its publications and impacts retrospectively become invalid, like in quantum entanglement. This, it seems to me, is not what the casual frequentist would have intended. The problem can be patched up by changing the criterion for validity, from whether the researchers don't know the value, to whether no value exists in theory, but that pulls down the edifice of inference on the population parameter π. An alternative escape route seems to be to move from from whether the researchers don't know the value to whether no human knows the value (or maybe *could* know it), but this is impossible to confirm, or infects frequentism with a strong dose of subjectivity. I have heard these arguments, but they misunderstand the foundation of frequentism (the long run) and are quite quickly dismissed. That people still think them indicates a lack of respect for foundations.

unknowns should be impossible.

But what about unknowns in between, a grey area including group-level effects in a multilevel regression? In some circumstances, for $n \leq \nu$, where $0 \leq \nu < \infty$, new draws from the aleatory sampling distribution appear, but sooner or later, the groups change. Frequentism traditionally forbids estimation or inference of these parameters for any n[†], but does (because it is useful, I suppose) allow estimation and inference on other parameters, including the variance of such "random effects" after the group-level parameters (or "effects" for sensitive ears) themselves have been integrated out of the joint likelihood.

Some frequentists might then calculate (in a two-step process to avoid contamination) BLUPs, which are effectively empirical Bayes posterior means and standard deviations. To complete the humiliation, these must be called predictions (the P in BLUP) not estimates. What a strange accommodation for the sake of realpolitik.

However, it seems to me that some subjects of study have quite different limits, ν, on aleatory re-sampling. If the groups in the multilevel model are students taking a course, we expect them all to be different in next year's data, so $\nu < 1$. When ecologists collect data along transects and the transects are the groups, we expect the transects to persist for a very long time. Perhaps they could in theory walk the transect several times a day, every day in the same season, maybe 1000 times a year, for maybe 1000 years or more, so $\nu \approx 1{,}000{,}000$ (let's assume that they have a way of adjusting for climate change). Transects persist, while students change. Why, then, should we not carry out frequentist inference on those group-level parameters? Surely it is consistent with long-run (not eternal-run) frequentism to consider ν and talk about it up front in justifying the choice of method. The fact that this does not happen is, for me, a symptom of the lack of philosophy among statisticians and quantitative researchers.

I like to clarify what the inference in a project is *to*: what the population is, how the sampling distribution emerges. When all the relevant data were present (as in complete administrative data in a retrospective project), I have several times refused to produce any inferential statistics, because there is no role for them: the aleatory sampling distribution is degenerate[‡] because repeated data collection will always return the same data. This irks collaborators who hold on to p-values and confidence intervals as performa-

[†]Why do so unless you really believe there is no probability except for in literally eternal runs? However big ν is, they want more. My wife and I were in a café in Pittenweem, a quaint village in Fife latterly populated with white-collar types who work in Edinburgh, and I found a copy of a research report on the village's arts festival lying around. The author maintained a straight face as they reported answers to the question *How long have you lived in Pittenweem?*: "some free-text answers could not be classified in the pie chart above, including 'longer than you' ".

[‡]A single point of non-zero probability in parameter space. Show-offs call this a Dirac distribution.

tive shibboleths. I don't blame them for this; it is what they were taught (see Section 9 for more on education).

Now, consider the canonical case of an election poll. Do we refuse inference because the election will happen soon and then repeated data collection becomes impossible? I think not, because of an assumption that has so far gone unquestioned: that these hypothetical repeated data collections must happen *sequentially*. Why? Why not simultaneously, in which case we can do inference any time except for when we have whole-population or "census" data. (It may be too provocative to ask whether repeated data collections going backward in time are admissible, and if not, why not.)

There is one more consideration here and that is time series, or more broadly, structured data. Suppose we have complete data on time points $(1, 2, 3, \ldots t)$ and a model. We must make predictions for $(t + 1, t + 2, \ldots)$. (The same reasoning applies if we have geo-located data and must predict other locations.) Our prediction involves the predicted value from the model, plus the estimated variation around predictions. I think it is correct to call that a prediction interval because it serves a different purpose to a confidence/credible interval. In such time series predictions, the "population" is all potential vectors of future samples, not resampling of the current data.

If we did not have complete data, we would also have aleatory uncertainty around the model parameters and hence the prediction; our prediction interval would include additional uncertainty from the combined effect of the joint sampling distribution of the model parameters; it would be a kind of hybrid confidence-prediction interval, but its purpose is prediction, so we call it that. I think this underlines the need to consider inference on a case-by-case basis.

Note that prediction with such structured or correlated data-generating processes is different to simple independent events, because new information on one event changes the predictions of some others. Suppose we have prediction distributions/intervals for $t + 1$ and $t + 2$. Then, the data for $t + 1$ arrive. This immediately changes our prediction for $t + 2$, and not simply via a change to the likelihood of the parameters, but because of the correlation / structure. (This accounts for the great rise in popularity of Bayesian Gaussian process models in recent years, in spite of their computational challenges.)

The outputs from Bayesian statistics, such as the probability that a parameter lies in a certain interval, are useful. It is also very flexible, especially because it is amenable to inference by simulation, so the asymptotic properties of estimators and standard errors need not be proven, and the pesky normalising constant can be ignored.[§]

[§]To be fair, you can use the same simulation methods, like Markov chain Monte Carlo (MCMC) on likelihood-based inference, it's just that nobody does (setting aside the case of flat priors, a kind of (mostly unintended) likelihood-based inference by default). I shall refrain from conjecture as to the reasons, but I suspect the blinkers of tribalism and

There are some problems with all of the methods listed here, and none of them is universally applicable. You can't get p-values from Bayes, for example, and you can't get posterior probabilities from frequentism. There are some work-arounds, like confidence intervals, which are only permitted in frequentism if they are given a lengthy and tortuous definition.¶ People have to choose one definition and stick to it, or so it seems. There are many who use casually eclectic methods, but are open to criticism for it. As I will explain in Sections 6 and 7, I think that the frequentist definition and the Bayesian (or other[5]) definitions can be equivalent.

There is one more thing I will point out here about the old arguments between frequentism and Bayes, a kind of semantic turf war. We may not call the likelihood a probability because it does not integrate to one over its support, which is an essential property of probabilities — fine. But there are other conventions in the names that we give that arose simply from frequentists (historically the vast majority of almost all statistics departments in universities in the mid 20th century) refusing to allow their terms to be used for methods that broke their rules.

So, we have a credible interval rather than a confidence interval, a posterior standard deviation rather than a standard error, and predictions of individual random effects in multilevel models rather than estimates. I don't like this sort of posturing very much, and I'll come back to it in Section 3. Interestingly, the concept of significance, though it is typically tied up with p-values, is not their sole preserve, for the reason that it is standard practice, even if it is highly ritualised, to pre-specify the meaning of significance (after Neyman and Pearson — for example, "significance was defined as $p < 0.05$"), so why not pre-specify it in terms of a posterior distribution?

---

performative science, of which we will hear more later.

¶ "A function mapping the data to an interval, which will, when applied to each of a long sequence of n identically repeated events, contain the true and unknowable population parameter θ a proportion of times $\hat{\alpha}$, where $\lim_{n\to\infty} \hat{\alpha} \to \alpha$". Thus, a staunch eternal frequentist might reject statistical inference of one-off events such as election polls, because $n \leq \nu \ll \infty$. This leads to a well-worn riposte that the finite life of the Sun or of the Universe renders all frequentist inference invalid. The fundamentalist frequentist must either demand literally eternal identical replications and fall foul of the second law of thermodynamics, or relax it and fall foul of a sorites paradox. But let's be good sports and focus on the utility of their methods, which work and give the statistician and their audience satisfaction by returning a well-behaved estimate $\hat{\theta}$ and confidence interval $(\widetilde{\theta}_l, \widetilde{\theta}_u)$. I'm also being a bit rough with frequentism here by attacking the infinite-run, which not every philosophically-well-informed frequentist would subscribe to. There are interpretations that use a long-run definition without requiring eternity (let's say they are aiming for $\nu^\star$), though they introduce other problems,[5] even if you achieve clarity on ν. Notably, there are propensity interpretations, where probability quantifies an attribute (human-defined, not innate) of a system, which produces frequencies that are more or less proximate depending on the "run". They step away from a direct realism, but they are vanishingly rare in practice. The ontology of frequentism and Bayes, the dominant interpretations, will be tackled below. It just seems ironic that frequentists traditionally mock Bayesian statistics as a religious cult.

3 Communication, performativity, postmodernism and humility

Lyotard described performative science as a culture in academia where the impact of research does not matter, but rather the demonstration of adhering to the rules of conduct: performing.[6] We may not have to try too hard to imagine a university department, the School of Moribund Studies perhaps, where pointless papers are published for the purpose of managerial targets, citing the right influential people's work favourably, and using fashionable methods and turns of phrase without regard for their utility. For Lyotard, this is a residual culture of modernism, protected from changing times within the ivory tower. There is no suggestion that they should write for anyone other than their own colleagues and rivals.

For me, Augé defines modernism and postmodernism in the most relevant way.[7] His modernism is the period in European / North American culture when a privileged class of people are permitted to practice science by observing others. An anthropologist, for example, after study and demonstrating their credentials by performance (in the Lyotard sense) in Paris or Boston, would visit and write at length about a tribe of people living in a rainforest.

There would be no suggestion that the anthropologist brought their own biases and cultural interpretations to bear on the observations: modernist science is positivist (there is only one true set of facts) and often assumes that humanity is on a Hegelian journey of inexorable progress. It was inconceivable that the people from the rainforest might have their own views and voices heard about their way of life. To suggest that they might visit Paris or Boston and comment on the ways of the inhabitants was only the stuff of comedy.

In humanities, this modernist approach has faded away, but in parts of science, it lives on. Unfortunately, I have seen the attitudes of the School of Moribund Studies appear in the statistical aspects of otherwise sensible, wise and caring researchers. This is how most p-hacking occurs, how data get shoehorned inappropriately into simple tests and stats. They are not malicious, but they have had minimal training in statistics, cannot find a statistician collaborator, and feel pressure to *perform* as they saw their professors do before them, by printing the trappings of statistics. As they say in the Cape, where I grew up, *basela ngendebe endala*, they drink from the old cup.

Interestingly, many of my former academic collaborators came from professions such as nursing, where a fight for professional identity and autonomy in the second half of the 20th century coincided with feminism, postmodernism and the so-called paradigm wars in research methods. The paradigm wars pitted adherents of quantitative methods against qualitative

researchers. This seems bizarre to most people now, as they use different data to achieve different, and generally complementary, goals. The qualitative v. quantitative conflict, and the frequentist v. all-comers conflict, both had strong elements of performance about them.

So, my collaborators would often value the qualitative aspects more than the quantitative. They might ask me to crank the handle and produce a p-value or two, but in many of these projects I found more interesting and subtle patterns in the data that warranted more complex, and often Bayesian, models. It was a fertile ground for both applied and methodological work, and we all learnt from the projects and from each other. In particular, I came to appreciate the value of understanding different views of the data and the reality it represented.

Hence, my work is postmodern. Rather than adopt a positivist stance or a social constructivist one, as many of my colleagues did, I was drawn more to complexity theory. This led to my paper with Rick Hood on trouble brewing within public health research.[8] Society, and large organisations like the National Health Service, are seen as complex systems. There are many interacting parts, and plenty of intelligent agents. They adapt to circumstances, and so the effect of an intervention can be highly unpredictable. A small input can sometimes snowball through the system into a large effect. Some patterns recur over time, but not in precisely the same way.

To research this sort of system requires, first and foremost, the voices and views of those who are within it. (I will talk about how we can do this in Section 5.) The privileged external observer will be doomed. We need humility, and a wide range of data, both qualitative and quantitative. Although statistics will help us predict what will happen, it will be imperfect. We must set aside the performative notion that an estimate or significance finding is an immutable aspect of reality (this sentence, itself performative, gets backed up later, but we have to get into some tangled thickets first). Humility is perhaps the most important single attribute.

4 Communication; quantitative and explanatory (abductive) inferences

Data analysis is always part of a larger process. The data come from somewhere, and as I described in Section 3, that needs to inform the analysis. The findings also go somewhere, and are used by a decision-maker. They are usually not statistical experts. It is helpful to meet the audience half way, so that our work may influence them and have impact. This is another reason to use Bayesian statistics, because of the intuitively interpretable outputs from $P(\theta|X)$.

We must work hard on communication, and study good practice from others, with just as much energy as we would study some minutiae of statis-

tical modelling and computing. If we do not communicate effectively, all our clever analyses are pointless. I have experimented with some different ways of bridging the gap between their understanding of those basic Snedecor stats and the Bayesian structural equation models that I was producing. Some were quite successful ideas, others not so, but it was all valuable learning. In short, I contend that we must, above all else, serve the audience. That will require some experimentation, conversation and compromise: not terms we often hear in statistics, at least the modernist conception of it, where the number cruncher knows all and remains aloof. I drew great encouragement for this from Monteiro's code of conduct and colourful talks, captured best in his book about being an ethical designer in the early 21st century.[9] Mentally replacing design with statistics, I find an almost perfect correspondence, and a lot of wisdom.

I also think that, when the occasion calls for frequentist outputs like null hypothesis significance testing, we should use it. We just have to make sure that the audience understand what it can and cannot tell them. The only practical problem I see is when quite different interpretations of probability rub shoulders in the same project. I will return to this eclecticism in Section 6.

The role of prior distributions, what they represent, and what influence they might exert on the posterior, are sources of anxiety among those who do not use Bayesian methods.[10] I have mostly used either diffuse or weakly informative priors, with an aim of aiding computation and penalising highly implausible parameter vectors, but no more influence on the posterior than that. * However, there are times when the project really is about modelling how attitudes and beliefs might be updated by data, and in such a case, a consensus subjective prior is useful.[11]

I have never used an individual subjective prior, and I cannot imagine a situation where I would. My objection is purely practical; the philosophy and mathematics are sound. Perhaps this arises from the early part of my career, analysing biomedical studies and presenting the results to committees of experienced healthcare professionals. I could obtain the opinion of a consultant cardiologist†, and update it with data, but when the time came to show them the posterior, they would doubtless want to think about its implications and apply the old Art Of Medicine. In fact, they should stop thinking, because the posterior *is* their new thought. They could, for example, be replaced by a computer randomly allocating patients to treatments

*Priors that are not explicitly informative of an individual or a group's beliefs act on the likelihood in just the same way as a penalty term, for example in LASSO or ridge regression, where modelling preferences — prejudices, to those who disagree — are also enforced by massaging the likelihood. Yet one is accepted in frequentist practice and the other is not. Below, I will show that this apparent paradox is one example of a misunderstanding about the true divisions between methods in the foundations of inference.

†With apologies for stereotyping; but you all know the sort.

on the basis of the posterior. (Actually, as Glymour pointed out, this is not innate to subjective probability but instead to viewing subjective probabilities as a *norm of belief*.[12]) I don't want to live in that kind of world, but it is nevertheless worth noting as an aside that a genuinely autonomous AI system can work effectively with updated subjective priors. However you intend to use these probabilistic inferences on belief, I agree that they have their place, that they are mathematically and philosophically sound, but just not that useful to me.

The priors that I use tend to address problems like biases. I face a communication choice: either I elicit exogenous information and try to incorporate it into a holistic assessment, or I stick to the likelihood and then leave the audience with a Limitations section telling them that there are such-and-such problems with the findings they have just read, problems too big for the statistician to handle, and it's up to them to deal with it. My experience has been that the audience knows far, far less about evidence, statistics and mathematics than me. I think it is my professional role to help them, not to wimp out at the last minute and leave them in the lurch.

I am also interested in the way that statistical inferences are often extended by the audience into an explanation of how the world works. This sounds like causal inference, and indeed there are occasions when causal methods would be helpful to our audience. But, there are also many times when the audience want, quite naturally, to take statistical results and extend them into a new hypothesis. Many of the projects I have been involved in were both inferential and exploratory, hypothesis-testing and hypothesis-generating at the same time.

Statistical tradition, mostly but not exclusively frequentist, might suggest that this is wrong, and that the hypothesis-generating should be left to the audience after the statistician has stated the facts and left the room[‡]. I think this would be doing the audience a disservice. I framed statistical work as inductive or deductive before, but there is a third category called abductive learning. Lipton wrote about this with the name "inference to the best explanation".[14] He characterised a best explanation (for observed phenomena) as one that is both Likely (it fits with the data, perhaps using probability) and Lovely (it provides the most additional understanding, perhaps by matching other external results or related phenomena).

This inference to the best explanation is not the same as causal inference;

[‡]The pioneer statists of Germany, then revolutionary France, were sometimes even more extreme, viewing any summary or cross-tabulation as overstepping their duty. Instead, thick almanacs with many fold-out pages of tables would be prepared for, and presumably ignored by, royalty and politicians.[13] The opposing view was the Anglo-Saxon "political arithmetic". A lesser opposition remains today: economists, for example, are habitually comfortable with boiling many inputs and assumptions down to a single conclusion, while evidence-based medicine tends to require each study to be presented as one independent fact in a constellation, the form of which only the Aesclepian initiate may discern.

it is about generating the next hypothesis in a somewhat systematic way. I see this as helping with Platt's "strong inference", where scientific practice (not the performative stuff) is a spiral, moving between inductive exploration and deductive intervention, steadily towards (or is it?) the truth.[15] Platt says the entire process is inductive. I now think that the hypothesis-generating parts within it are abductive (to the best explanation, approximately). There is still plenty to be done to advance Lipton's work, but the framework helps us to talk about these next steps with our audiences.

5 Complex models rendered identifiable through constraints from exogenous theory

In order to achieve a complex, flexible model that takes into account the contextual, almost qualitative, information from human beings involved in the system under investigation, I favour Bayesian latent variable models. Before the advent of probabilistic programming languages like BUGS and Stan, there were only very limited tools to fit such models. A latent variable is a vector-valued parameter (or higher-dimensional, matrices and tensors) which varies in some way over the data. Random effects in multilevel models are latent variables, and so are imputed missing data. But some of the most valuable applications come when we impose some internal structure, some theory-informed constraint in other words, on the way the latent variables determine the observed or manifest variables. This involves talking to experts and those involved in the system under investigation to find ways to constrain the model.

Structural equation models (SEM) are an extension of the latent variable idea, where there are multiple latent variables, some of which (linked deterministically to manifest inputs) cause changes in others (linked to manifest outputs). A recurring challenge is to constrain these models so that they are identifiable, i.e. a single set of parameter values can be found that maximise the likelihood / posterior. The specification of the SEM may be enough, or other simplifications may have to be agreed with the stakeholders, including severing connections among variables, and imposing artificially strict prior distributions.

A realistically complex model like this (to borrow an apt term from Goldstein and colleagues[16]) requires rigorous evaluation. It is all too easy to optimise a model which is itself not fit for purpose, without questioning the model itself. The Bayesian workflow promoted by Gelman and colleagues helps here.[17] Posterior and prior predictive testing are falsificationist in outlook, tenuously similar to hypothesis tests. From the prior distributions, we generate a draw of parameter values, then plug these into the likelihood to generate a draw of pseudo-data. We repeat this many times, and compare the result to the observed data, which should be contained comfortably

inside the pseudo-data distribution. In the posterior version, we are sampling new pseudo-data from the posterior contingent on the data (which are fixed) and the parameters (which are not). If there is a mismatch, it may indicate that something has gone wrong with the sampling algorithm, or, more usefully, with our model specification. Like Feynman, we should bend over backward to prove ourselves wrong. Again, we serve the audience above all other considerations.

6 To justified eclecticism, via ontology

Before proceeding to questions of whether we can use frequentist methods sometimes, and Bayes at other times, I must address the question of the meaning of probability in my own Bayesian practice. I already said that each of the interpretations of probability can be used without undermining the mathematical validity of what follows. My view, in agreement with Spiegelhalter, is that there is "no such thing as probability".[18] That is, it is only a mathematical tool, and arguing over whether it may be used for epistemic uncertainty is similar to computer scientists arguing over whether one integer may be divided by another: an argument about using technology that says more about tribalism and performance in science than it does about evidence.

6.1 Thing one and thing two

My two concerns about staunch frequentism are the infinite replication and the identifiability of a reference class. Realism of an infinite collection of identical replications is probably only insisted on by people who have not understood the arguments. We dismissed it earlier. But even those who see long runs converging to an asymptote as the only definition of probability make an appeal to realism: that the things being counted exist and so there is no space for opinion, belief, &c (to do otherwise must therefore be Bayesian buffoonery). In other words, everyone who sets out to count these things will come to the same answer.

You recall from earlier that frequentism cannot be about who knows or might know, or even conceivably could know, the total count of an infinite reference class, because the damn thing is infinite. Now we must deal with the suggestion of a large but finite and countable reference class or real things as the basis for the only true probability. (Even Fisher used a finite reference class in his famous "exact" test.)

To what extent does frequentism actually rely on a realist interpretation of probability? Philosophically-minded statisticians and statistically-minded philosophers hardly talk about this at all. A conference on "Ontology and Methodology" in 2013 apparently consisted of talks about other subjects,

more concerned with empirical effectiveness of methods than realism or its alternatives, according to one attendee.[19,20]

Glymour pointed out that realist and antirealist standpoints on theory and reality both have problems.[21] They seem to me to point to different subjects, the realists to physics and the antirealists to social sciences, for example. The latter are more like complex systems, aggregations of many interacting influences. So, I think that there may well be real classes of real things, but those are perhaps limited to the molecular scale and below. I will call these *simple things*. There are also "things" that are more complex and fuzzy around the edges. For example, molecules of ibuprofen are real in this sense, and their isomers are a real property, perhaps because they are also real, with R+ibuprofen[†] simply being a real subclass of the class of things called ibuprofen molecules[‡].

This offers an alternative approach. Realism holds for simple, more fundamental, things, but starts to break down as we try to identify and count larger aggregations and interactions of those simple things. An example of a complex thing that everyone can, hopefully, agree on might be the effect on health of a tax on sugary drinks. Different theories will lead to many different ways of counting this. Horgan and Potrč, whom I will introduce later in this section, call the simple things snobjects, and the fuzzy aggregate concepts slobjects.

I often return to Boyle's law of the expansion of gases as my archetype of a physical law. It arises from the accumulated simple interactions of simple things, and any noise (influences other than temperature, volume and pressure, for example radioactive decay) that raises the movement of this molecule or lowers that one will cancel out long before there are 6×10^{23} of them. As we measure gas pressure in high school, we find it sticks to the predicted curve pretty well, although there is a little measurement error along the way. There is no need for statistics, but more complex settings involve a lot of noise, which we include as a probabilistic element in a model. The model thus has a deterministic element and a probabilistic (stochastic) one: the scatter around the curve. In complex systems, the problem gets even worse and the deterministic curve is highly unreliable over time too (remember Google Flu Trends?[22]), unlike Boyle's law.

[†]R+ and S- ibuprofen are mirror images of each other. R+ is an analgesic drug (and sometimes gut irritant) by binding to certain receptors in the human body. S- does not match the shape of the receptor, just as your left hand does not fit in your right glove, so it is medically inert.

[‡]If you have read about the philosophical subject called metaphysics, you might object here to my crude handling of properties and of the persistence of identity. Bear with me for a few more subsections.

6.2 Probability and reference classes

What are the implications for probability? I contend that probabilities are not real properties of real things but the cumulative manifestation of all the properties on all the things that have some interaction in this system. We represent them as probability because we don't care about all the minutiae of the system, and generally we couldn't evaluate all the connections even if we wanted to, which is also what we usually mean by that strange word *random*. (The notable exception to this is quantum probability, which is a real property in the same sense as the isomer of a molecule. But that is a quite different thing, mathematically and physically.)

And yet, this Spiegelhalter-style probability still manifests as an asymptotic frequency over long runs. Like any proportion, it is a numerator divided by a denominator. both being natural numbers. The denominator is the *reference class* and the numerator is the number of times that it matches some criterion. The frequentist realist argument is that a long run of identically duplicated experiments provides (or could provide) these two numbers, while anything else could never provide them*. These numbers must be at least capable of referring to real things, therefore only frequentism is possible.

If we are not going to insist on realism, then we are taking a metaphysical stance called nominalism. This asserts that classes of things do not really exist, but are just a semantic convenience. The important question for any statistician who is still in agreement with me at this point is how they will justify nominalism. However you get there, the rest will follow (in the next section).

Walking across the fields you see a bird; it is a magpie. It is a thing that belongs to the class "magpies". You get closer and find it is dead. Still a magpie? Some months pass and every time you go to the farm shop, you see this mouldering bird. Still a magpie? Some muddy matted feathers and bones — still a magpie? Soil — still a magpie? A hawthorn sapling grows from the spot. When did it stop being a magpie, when did it leave the class?

And also:

DAD: There are no quaggas any more.

DAUGHTER: Yes there are, stuffed in museums, ha ha.

DAD: That's not what I meant.

(By the way, if you like this kind of thing, you'll enjoy Section 7.) Could it be that the class of magpies and the class of quaggas are just fuzzy semantic conveniences (FSCs)? I'll try to firm this up, and look into all the potential counter-arguments, from here until the end of Section 6.4. Remember that

*A victim of performative staunch frequentist indoctrination might argue here that assessing degree of belief is not compatible with long runs, because the first experiment changes the researcher's belief, even before $n = 2$. But who said anything about sequential replications? Remember, we can run them simultaneously, perhaps with identically replicated researchers too.

it is a philosophical issue, so there will not be one universally proven answer, just one that I have settled on.

Now for statistical inference. Obviously, when people say "identically replicated experiment", they don't really mean it. It has to generate new data, so it has to be identical except for the sampling. And remember, the collection of not-exactly-identical experiments that we dream up must be really feasible, or inference will be forbidden. Complexity blows that out of the water: minute differences in experiment will potentially lead to big differences in data, and change the system, which matters even for simultaneous data collection. It's not clear whether we can even do it for the simplest things, as demonstrated by recent work on quantum computing. I won't even mention collapsing wave functions and such.

So, how far from the original experimental setup is OK? At some point, we will go too far, leave the reference class and not be able to count that replicate in our realist long-run frequency. To be a realist frequentist, one must define that boundary, not by personal belief (heaven forbid!) but by some attribute that makes a replication experiment incontrovertibly in the class or out of the class.

I am not denying that the asymptote exists as a mathematical construct, nor that it is important to evaluate it and use it for inference. I just don't see the relevance of whether or not it can be *empirically* reached by counting long-run replications. A temperature of absolute zero is an asymptote; as best we understand things, it does not actually occur anywhere in the universe, nor can it actually be realised, but we can calculate where it is from observing the curve that approaches it. Would anyone really claim that physicists must not talk about absolute zero or use Kelvins as a unit of temperature for that reason?

I understand questions about how close one's method can get to the asymptote; indeed, this is what statistical methodologists spend a lot of effort assessing with bias, coverage and efficiency.

If we reject the frequentist realist argument that probability *is* one thing and not another because of a real reference class (and we don't have to reject the idea of long-run proportions to do this), then we see that the confidence interval is *both* the proportion of times that the function catches the true value, *and* the chance that the true value is inside the current CI in front of us. Under any given set of circumstances, these two statements are either both true or both false; we will return to this in Section 7.

Propensities, themselves mathematical models of an asymptote, are quite different: they do not need to appeal to long-runs or real things, but they do need a reference class.

Simple things which are real can be counted. If a statistical model tells me the probability that the proportion of R+ ibuprofen in a batch might fall below a purity threshold, then we can count the molecules in each batch, classify the batches and verify the claim in the long-run (at least in theory).

Probability is here an accumulation of simple, real things (which have a reference class) and real properties (subclasses).

If it tells me the probability that children's mean body mass index will fall after introducing sugary drink tax by more than some minimum important difference, verification gets tricky. Maybe you get the data from school nurse visits. What if some parents, irate at this mass government manipulation, refuse to consent to have their kids measured? What if their kids are not like consenters' kids? What if the schools where you do the study (because you surely can't be funded to do it everywhere!), seeking good publicity, push for healthy diets in other ways too, because they know they are being watched?

Those are biases, but there are causal complications too. You might not be counting what you think you are counting, or it might not be *about* the same thing. Either way, if you repeat the experiment, the whole complex system will have shifted (even if you do them simultaneously!) and you will get results that do not come from a sampling distribution with one consistent reference class.

What if they buy counterfeit, tax-free drinks in a corner shop* that ignores the law? What if drinking expensive, high-sugar drinks has become a macho status symbol among teenage boys? What if they spend their money on chocolate bars instead? What if they save money by making their own sugary drinks at home? Why are you not counting urinary tract infections or levels of concentration in classrooms or eating disorders? Is obesity really the target, or should it be cardiovascular disease, arthritis and Type 2 diabetes, fifty years from now? Was it the tax or the publicity around it? Or did the tax and the weight loss both stem from changing attitudes, so it would have changed anyway? On and on the fuzzy boundaries of the reference class go.

That is complexity in action. And yet these kids, their drinks, the nurse and the weighing scales are all made of simple things, which just happen to have accumulated their interactions in a complex way, and probability can no longer meaningfully appeal to a reference class.†

It would make (more) sense to use frequentist methods on the ibuprofen purity analysis, and (less) sense on the sugar tax analysis. This is because of the structure of the network of interactions among real things (those that have a reference class). Some networks acquire complexity along the way, and in so doing, lose a single reference class that can be applied to repetition

*https://www.gov.uk/government/publications/annual-ip-crime-and-enforcement-report-2019-to-2020/ip-crime-and-enforcement-report-2019-to-2020

†It does not help to measure something very simple as a proxy, or just use someone else's data which is sitting around (fuzziness? what fuzziness?), extrapolate to the reference class you would have liked to have measured using someone else's model which is sitting around (model space? what model space?), and pretend that you have done the job properly.[23] This sort of thing is increasingly accepted in performance; we all have to pay the bills.

under nearly identical circumstances.

6.3 As one with the cosmos

How do we justify the ontological switcheroo? How can we be sure that I am not just switching when it suits me? In short, by relegating even the simple reference classes to the status of semantic conveniences (slobjects), aleatory and epistemic uncertainty are united in miserably fumbling in the dark for answers. Sounds about right to a practising statistician. This is a nominalist view of the metaphysics of attributes and classes, which I will explain shortly. I'll set out how I arrive there in this subsection, then consider other ways in the next.

This is the part where you are most likely to conclude that I have "gone wacky", so let me try to be as brief as possible without getting cryptic. The universe is made of matter and energy, which appear and disappear through various forms. It is more helpful to think of it as one convoluted four-dimensional object in space-time*. Our human bodies are no different; bits are added and removed all the time. It seems to me that *there are no objects* when we get down to fundamental facts, only patterns that are sustained, morph and reappear in similar, synecdochal form over space-time (the parallel to the meta-stable patterns† in complex systems is not a coincidence). Objects can only be defined, somewhat subjectively, by human language, which places them in a reference class, the most trivial being "this object". So, the only real object is the whole Universe. The "things" I referred to earlier are actually just meta-stable patterns. This makes me an *existence monist*, which is a rather unusual view.

Attributes of anything are attributes of the universe, because of existence monism. There is only one "sequence of events" — in fact it is one complicated object in space-time — so all long-runs and all beliefs about unknowns are long-runs of the same object and beliefs about the same object. The only difference is how we define the criteria that first admit certain meta-stable patterns to the reference class and then a subset to the numerator — we chop up the universe into questions of interest and get probabilities for those. Without a trustworthy reference class there can be no meaningful

*Here, and in what follows, I subscribe to, and rely on, the "block theory" or B-theory of time, that matter and energy exist across space-time, even in that region we call the future.[24] However, there is also a "growing block theory", which sees the past and present as a physical 4-D object, but the future as a growing edge. Theories of time are among the hardest philosophical concepts for which to find a rationale. I just find the B-theory consistent with the rest of this section (which I admit is a flimsy rationale). Below, I give a rationale also involving semantic attributes versus reality, which I *think* extends my conclusions to the growing block theory of time.

†"Meta-stable configurations" is a phrase I quite like, from a paper by Itay Shani. But I prefer patterns, as configurations must be configurations of something, and there are no objects to be configured.

counting of long-run frequencies under nearly-identical repetition.

The important idea here is that *all* reference classes are just linguistic constructions (a model) applied to meta-stable patterns in one object (Horgan and Potrč call it the "blobject"[25]). Even the ibuprofen molecule we encountered earlier arises out of atoms, arising out of other atoms and subatomic particles inside a star, arising out of quark-gluon soup, arising out of the post-Big-Bang something-or-other. It is therefore one pattern in a space-time swirl of bits arriving and leaving, just like you and me, although it has fewer simple thing parts.

And it is semantic: suppose we make one of its carbon atoms the radioactive isotope carbon-14. Is it still ibuprofen? Yes. Why? Because I say so. Someone else might disagree. If this makes me an antirealist of probability, then remember that it is down to an ontology of probability rather than an instinctive suspicion of theory or social constructivism (the usual route), so don't put words in my mouth or associate me with hippies or Left Bank poseurs.

You might reply that however complicated it is, we can still count parts; it may be hard but not impossible. I don't deny counting, I just see that two rational observers will come to different answers because of the growing complexity of the network of things and hence fuzziness of the reference class. You can't agree on a count of things if you don't agree, now and forever, on what a thing is. As a self-indulgent final attack on possible exhaustive counting, there is the problem of thermodynamics that we met before: we can't, for all reference classes, count all parts across the time dimension. Also, the problem of black holes: we can't, for all reference classes, count all parts across the space dimensions either. It's not looking too good for possible exhaustive counting, or for realist reference classes: neither numerator nor denominator, neither Thing One nor Thing Two[‡].

Reference class names are attributes but not things, nor properties because they do not make subclasses of things. (The meta-stable patterns also are not things.) In turn, the attributes are not intrinsic but a subjectively defined description. The frequency follows directly from the numerator and denominator, but belief about the meta-stable patterns *is the same*: an extrinsic attribute or description leading to a number between 0 and 1.

Now, consider missing data, an example of "epistemic uncertainty". The datum is there in the four-dimensional universe but not available to the part which seeks to investigate it (us)[§]. Likewise, "aleatory uncertainty"

[‡]From *The Cat In The Hat*, because ontological things also seem innocuous but go on to cause a lot of trouble.

[§]This necessitates a standpoint on the philosophy of consciousness too, because awareness and memory are entailed in some of the interpretations of probability. The Loveliest (as Lipton might have said) explanation seems to me to be the cosmopsychism most closely matched by Goff[26] (except that he is not such a strict existence monist as me, and hence often refers to his thesis as panpsychism) when he writes that consciousness is an aspect of

represents meta-stable patterns extending into the future, and hence also not available to the part which seeks to investigate it[¶]. I conclude that without realist interpretation of probability, there is no distinction between aleatory and epistemic uncertainty, and so belief about the universe (the parts that are not "us" appear to "us" as a data-generating process: DGP) is simply a different estimand to frequencies in the universe (arising from the DGP).

If we cannot rely on realist reference classes, what are we doing with probability? I asserted that probability functions are mathematical models for the aggregate behaviour of a network of simple things. We have imperfect classification for things and properties, so those networks are hard to define. With few interacting parts, we are on very firm (though never perfect) ground, but as the network grows or complexity kicks in, it becomes very shaky indeed. However, there are not two kinds of probability, just a sliding scale of discomfort for the analyst. We just come back to having a model of how the world works, a mathematical simplification, which, we hope, will help us to understand it better. It's not real, and neither is the simplest kind of model, which just counts things of a certain class. The universe is real all right, but we are not reliably able to chop it up into reference classes. Still, we can get numbers that are much the same most of the time if we are careful: a small amount of reference class uncertainty over and above aleatory and epistemic uncertainty.

Let's return to my earlier criticism of frequentists who do not have a clear

---

the universe, like electromagnetism, and given its propensity to pop up from nothing, must extend as a field through space-time. Again, I admit that this is a flimsy rationale. So, at the risk of being dismissed out of hand by those more committed to performativity, given cosmopsychism plus existence monism, concerns about the suitability of probability for epistemic uncertainty seem no longer defensible (unless one rejects probability completely (minus the quantum version, which is not a *meta*physical matter), which is an option but I don't find it useful because I still have to learn about the world around us and use numbers to do so), because there is no difference between aleatory and epistemic, in turn because there is only one object, the whole Universe, and one consciousness extends throughout it. (Like Criswell, "I" know all. "I" just can't recall it at present.) However, it is important here to consider the implications of existence monism (which I feel quite certain about) *minus* cosmopsychism (which is more speculative). The awareness and memory, which allow sufficient intelligence (artificial or natural) to define criteria for a reference class, which in turns divides a long-run of events (which, recall, already exist as *meta*-stable patterns in space-time) into binary categories and hence a frequency, subsists in the brain (or computer), and is unaffected by whether the underlying consciousness is a field or a collection of instances. It could be said, in fact, that my critical choice in rejecting realist aspects of definitions of probability is about time and ontology more than consciousness, but I leave it here for completeness and because it is the fastest route out of these gnarly bits of philosophy and back to implications for statistical practice.

[¶]The growing block theory of time would include the missing datum or bias in data collection as part of the monad, but not the future events in the reference class that help with a *realist* definition of long-run frequency, but we have just seen that membership of a reference class is a semantic, extrinsic attribute of reality, via existence of the reference class. Without the future being an object, there is only the semantics left.

idea of the long-run stability of the reference class, hence the aleatory sampling distribution, ν, nor how long a run they really require for hypothetical proportions to become "probabilities", $\nu^\star$. The purpose of evaluating your method "in the long run" is precisely to make your conclusions robust and reliable by acknowledging, quantifying and incorporating uncertainty. If single point conclusions are questionable, you need to expose that by giving yourself "a run for your money"[‖]. This, and any other inference, should improve the overall error rate of the entire data-driven decision-making process. I have no objection to aleatory-only inference, using the long-run, but I do believe that ν and $\nu^\star$ need to be discussed case-by-case.

6.4 Other ways to nominalism

However I might get to nominalism on the reference class of the long run, I believe my justified eclecticism follows. My principal route is existence monism, but there are other routes that do not go through such an unfamiliar and frightening forest, notably a stance called *resemblance nominalism*.

Philosophers often approach realism vs. nominalism from the field called metaphysics. A common subdivision of realism is over whether properties (reference classes are the case that interests us, as they permit frequentism; no other interpretation of probability is holding out against eclecticism in this way) exist in the objects that exhibit them (and hence are shared across multiple objects; this is *in re* realism), or exist independently in some abstract but real sense (are hence are somehow allocated by a kind of lookup table (presumably only viewed with light invisible, hid from our eyes) to the objects themselves; this is *ante rem* realism).

If there really is such a thing as the reference class of magpies, then there must be a property of magpiety, so where is it? *Ante rem* realism is tolerated in metaphysics because they also want to solve puzzles like where numbers are, and what they are. It might sound irrelevant to statisticians but I had to look down all these avenues in order to be sure that I had not missed a good anti-eclectic argument.

I do not care whether there are abstract objects or not, and I am rather suspicious of the confusion sown by that word "are" in such discussions. However, on the subject of particulars and universals I hold the opinion that all particulars are fuzzy and ultimately semantic conveniences, and building on this, I conclude that all universals are even fuzzier and even more limited to semantic convenience. That, I think, does for *ante rem* in this context. My view on FSCs is not one that I can turn back from from without becoming fundamentally a different person. You turn if you want to.

Now I must consider *in re* realism as pertaining to probability. Suppose

[‖]From Gil Scott-Heron

that magpiety is present in the aforementioned dead magpie, but leaves it at some stage of decomposition. We still have the problem that our statistical work must be mediated by the judgements of human observers, both data collectors and interpreters for decision making. The measure of whether we can use reference classes in a certain setting is whether rational independent observers (a phrase from Tony O'Hagan and Jeremy Oakley, in a different setting) would agree on the numerator and denominator that arise from it. When they don't, it is a FSC*. And I still contend that it is not a binary matter but a sliding scale of discomfort. So, even if there are *in re* properties, they do not seem to help us in this setting.

Of the nominalist standpoints, I favour resemblance nominalism, which says that these black and white birds share a name simply because they are similar to each other.[27] I like Nelson Goodman's definition of resemblance nominalism: resembling each other is what makes F-particulars have the property F, rather than F-particulars resembling each other because they have the common property F†. It is essentially a clustering task, and statisticians know that almost every clustering job leads to compelling labels that have no basis in reality‡.

Resemblance nominalism was, for a long time, regarded as a historical curiosity, but was recently revived by Rodriguez-Pereyra in an extensive reworking of the arguments. He relies on a multiple possible-worlds metaphysics, but in our statistical setting, I think that, if you regard not only the classes but the objects themselves as mere FSCs, then there is no reason why they might not also have fuzzy, semantic, and convenient, comparisons among them, the whole edifice being a great construction of words that often helps us make sense of the world. There is no need for realist possible-worlds.

6.5 Justified eclecticism

If aleatory and epistemic uncertainty are really two ways of looking at the same thing, then this suggests an eclectic approach. Bayesianism allows both; everything else — long-run frequencies, propensities, logics — restrict you to aleatory, supposedly because the reference class can be defined with sharp borders, but this is not the case. That is not to say that Bayes is the only way to infer. We can choose different meanings for different purposes.

In fact, the increase in the use of Bayesian methods in recent years has

*You might complain here that I am letting man be the measure of all things, and so begging the question, but think it is a more thoroughgoing problem than just the irreproducibility of *decisions* on the admissibility of repeated nearly identical data collections. If there is any scope for unforeseen disagreements on admissibility, then the whole idea of a long-run proportion is built on something not wholly objective and fixed. I cannot imagine anyone making a sensible argument that there is never any such scope.

†The choice of the letter F is Goodman's, but of course I enjoy the suggestion of Fuzziness

‡*viz* goo.gl/PSHNty

been driven in large part by eclecticism, where hypothesis tests are used for some tasks, maximum likelihood methods for others, and perhaps Bayesian or machine learning methods for specific challenges.[28]

However, Mayo (one of the aforementioned very clever people who disagree with me) is critical of this use of statistical methods without a clear philosophical underpinning.[29] I agree that it is unhelpful to be casually eclectic, not least of all because a mixture of findings from different statistical paradigms leaves the reader uncertain of how to interpret them. We see this a lot, for example, in network meta-analyses where pairwise comparisons are done frequentist-style, and the combined analysis Bayesian-style. The problem is that such casually eclectic analysis involves mismatched philosophical foundations, so the outputs cannot be contemplated together. And a holistic foundation is why we're here.

In response to Mayo's book "Statistical Inference as Severe Testing",[29] I set out my approach to using different statistical methods for different tasks in the table below. Rather than dividing tasks into those that include epistemic uncertainty (numbers unknown because we do not know them, like missing data) or are limited to aleatory uncertainty (random events from long runs, like sampling error), I reject the distinction for reasons given above.

Instead, we must be concerned with whether we are modelling just a data-generating process (DGP, which might include flat, diffuse or weakly informative priors; for brevity, and not without a little discomfort, I will just call these *uninformative*) or beliefs about a data-generating process (which involves more informative or subjective priors).

It is important that I note at this point that belief about the DGP does not necessarily mean whimsy, it is a means of bravely attempting to quantify exogenous information alongside the data and likelihood, for example unsystematically (tacitly) collected insights about a particular bias affecting the data. The numbers we will get back from our analyses are likely to be different, and will be interpreted differently, because there are different *estimands*.

Against this, I divide tasks into those that evaluate the likelihood or posterior at pre-specified parameter values (*a priori*, like hypothesis tests do), or at parameter values determined by the data and model (*a posteriori*, like maximum likelihood estimation does). Equivalently, we could say that we have to identify a parameter value or values, and also evaluate the likelihood / posterior at those values; the question here is which of these comes first. This, I think, is the crucial division in statistical practice, rather than the semi-mythical battle between frequentists and Bayesians. I believe that this provides a *justified eclecticism* where each of these methods is valid for the correct combination.

	Model of DGP	**Model of belief about DGP**
Identify then evaluate (*a priori*)	Neyman-Pearson tests, associated confidence intervals, Bayes factors with uninformative priors	Bayes factors with informative priors
Evaluate then identify (*a posteriori*)	Maximum likelihood estimation, associated confidence intervals; Markov Chain Monte Carlo, Approximate Bayesian Computation, *vel sim.*, with uninformative priors	Markov Chain Monte Carlo, Approximate Bayesian Computation, *vel sim.*, with informative priors

In fact, the *a priori* testing methods of Neyman-Pearson tests and Bayes factors also get used for *a posteriori* parameter values. An example is when we fit a regression and look at the p-values for the maximum likelihood estimates of the slope parameters (implictly, compared to zero via their marginal asymptotic sampling distributions). In such instances, the N-P tests should really be given some other name, because Neyman and Pearson the younger, wisely, did not advocate such a fast and loose use of testing.

In more practical terms, they tell us very little. Fisher intended them to be interpreted thus, as one little bit of evidence among many, by skilled and dispassionate observers, but the world is not as simple and well-behaved (and modernist) as he might have wished, and nowadays we see a lot of angst about such tests. I set them aside here; I prefer not to use them, but sometimes I compromise.

To be clear, I have no objection to a proper Neyman-Pearson decision-theoretic test. But Robert, you ask with a gleam in your eye, what about having no p-value, just significance compared to a critical value? I'll leave that up to my future assessment of individual projects, collaborators, clients and audiences. It can be dangerous territory, full of misunderstandings. I've done it, and I've also criticised others for doing it: fickle and changeable. I would rather elicit an *a priori* minimum important difference, and from a posterior distribution report the probability of exceeding it. That's just me and the sort of people I've worked with, though. It doesn't change the table.

I will conclude this densest part of the paper with a note that, although I have described procedures like hypothesis tests and confidence intervals in relation to the "true value" θ, this is a shorthand. The ontological rejection of probability as a real thing (and bearing in mind the complexity argument) also rejects the existence of θ. There is no true value, but perhaps a lot of slightly different values over space and time. Nevertheless, we can talk about

it as a kind of population-averaged value, in the same way that we can talk about long-run frequencies, subject to the challenges of the reference class boundaries.

7 Aboutness

I also find Yablo's work on "aboutness" useful.[30] He described linguistic and probabilistic misunderstandings in terms of the *ways* in which statements can be true or false, which helps to define what the statement is *about*. (I will call these Yablo ways, to avoid confusion arising from such a common word.) For Yablo (and I agree), two statements are about the same thing if they satisfy two conditions: firstly, that under any given circumstances, they are either both true or both false, and secondly, that under any given circumstances, they are either both true in the same Yablo way, or both false in the same Yablo way.

7.1 Aboutness and estimands

Let's return to the competing interpretations of confidence intervals, a function $g(\cdot)$:

$$g : \boldsymbol{X}, h \mapsto (\widetilde{\theta}_l, \widetilde{\theta}_u)$$
$$\text{where } h : \boldsymbol{X}, \boldsymbol{\phi} \mapsto \hat{\theta}$$

$\boldsymbol{\phi}$ are some nuisance parameter(s). I contended previously that they are either both true or both false for any given circumstances. By circumstances, I mean a true parameter value θ. We don't have to know or even estimate θ for this to hold. Now, this depends on stepping away from realist frequentism and eternal identical replications, but Yablo's second condition is more self-evident. They are made true or false by the same Yablo way, namely whether $\widetilde{\theta}_l \leq \theta \leq \widetilde{\theta}_u \,, \quad \forall \theta \in \{\boldsymbol{\Theta}\}$. If there is a certain probability of this being true for one sample $\boldsymbol{X}$, then that is also the long-run proportion, and vice versa.

This might be approximately true (Yablo's work includes the meaning of partial truth) under weakly informative priors or small-n MLE, but it will not hold for subjective, informative priors. So, I conclude that the frequentist and Bayesian (and other[5]) interpretations are *about* the same thing if the estimand is the same (the DGP or belief about it). Therefore, the critical division between columns of the table is supported: that *the estimand leads to different meanings and not the definition of a reference class.*

7.2 Aboutness and confusion by omitted ways

Yablo identified stereotypical and marginal Yablo ways, the stereotypical being more familiar to the audience (we might include cognitive biases here)

and the marginal less familiar. When the analyst and the audience have different Yablo ways in mind, confusion results like this:

ANALYST: These are the stats for ice cream sales, and these are for road traffic accidents, and their Pearson correlation coefficient is 0.6.

AUDIENCE: Wow! So if we ban ice cream, accidents will go down.

ANALYST: That's not what this is *about*. We want to use live ice cream sale data to predict accidents before they happen.

This is the most obvious confusion of Yablo ways in statistics: when the audience interpret correlation as causation. Causality is a very persuasive Yablo way, one that we seek out from infancy. Causation is indeed one possible meaning from a correlation, prediction is another, and the set of Yablo ways held by the audience here is not the same as the analyst's. (OK, they have both missed a glaring confounder, but it's just an illustration.)

Statisticians are usually careful to distinguish the Yablo ways that are closed off or still open to consideration, but sometimes in terms not understood by a non-statistical audience. Even when we translate, those parts of publications and discussions are sometimes omitted as intrinsically less important than the "bottom line" or the context of the applied field of study. In a way, they *are* less important — they only seek to serve the research process rather than the audience — until their omission leads to misinterpretation and undermines the value of the research effort. An example might be the assumptions of a parametric model, where a significant result might be true via two Yablo ways: that the population parameter is not the null value, or that an assumption is violated.

Omission of Yablo ways can lead to what Manski has criticised as "incredible certitude".[31] ("Ranges are for cattle! Give me a number.*") Certainly, more reductive models (those that rely heavily on *ceteris paribus*, give population averages as a confusing proxy for functions of individual circumstances, or neglect the adaptive nature of the complex system) introduce Yablo ways that are further removed from the realism of the data-generating process, which are likely to be "marginal" and hence overlooked.

In the paper with Rick Hood,[8] we quote an interview given by the erstwhile English Chief Medical Officer: "We know that sugar tax works because there is good evidence from Mexico". While the evidence from Mexico is convincing that sugary drink consumption fell after introducing a tax,[32] it actually says nothing about what "works". That is not what the Mexican analysis is *about*. There are many Yablo ways, such as increased consumption of *agua fresca*, which might be distinctly marginal to an English doctor.

This seems to leave a gap between statistical analysis and decision-making, which can easily be filled with cognitive biases. This is the space that Lipton's inference to the best explanation explores. From this, I draw more support both for flexible models and for communication that puts the

*Lyndon Johnson, to a room full of economic advisors, perhaps apocryphally.

audience first.

7.3 Aboutness and possible-worlds

Aboutness matters for the fundamental and semi-mythical difference between Bayes and frequentism. If theta never changes, there is no frequentist probability for it because such a probability is only a possible-worlds frequency. Each to their own. But the confidence interval and credible interval are more subtle. $P(\ell \leq \theta \leq u)$ can be seen as *about* θ or *about* (ℓ, u). It doesn't intrinsically differentiate them. Then this leads to my previous statement about how, in any given possible world with the same theta but different (ℓ, u), they have the same truth/falsehood and the same ways, so are equivalent (*modulo* approximations to the likelihood).

Yablo gives an explanation on p.45: "Aboutness is preserved [between A and B] if worlds where B is true in different ways cannot have A true in the same way".[30]

This made me a little concerned that I am equating the Yablo ways of the posterior distribution (DGP parameter), and therefore the credible interval, with the long-run frequency's confidence interval simply because by their definitions there *is* only one way for each to be true: θ lies between ℓ and u.

How could the CI/CrI be true except by containing θ? The traditional discourse has focused on the probability's arguments, whether it is $\theta|\boldsymbol{X}$ or $\boldsymbol{X}|\theta$. I move away from this by expanding the probability into possible-worlds. We must do this if we are seeking to define probability, or we will be begging the question. (And thus I dismiss propensity interpretations, hopefully not with a brevity suggesting I am as mistaken as Fisher's infamous "the theory of inverse probability is founded upon an error, and must be wholly rejected".) The new problem that arises is that we might argue over whether the Yablo way "$\ell \leq \theta \leq u$" is achieved via the movement of θ or of (ℓ, u). However, there is no movement.

In possible worlds in the frequentist definition, (ℓ, u) takes different values while θ does not; that is clear-cut. The Bayesian also expects (ℓ, u) to change with new data, and to shrink as data accumulate. The difference is that the Bayesian does not design (ℓ, u) in order to be true in a certain proportion of possible worlds.

(Possible worlds has not been appealed to in statistics, except more recently in counterfactuals, but that is different. Counterfactual worlds have different probability distributions, while replications of the data collection lead to possible worlds that do not contain probability until we count them all up together.)

Frequentists are innately sceptical about probability. They construct a definition out of possible-worlds, which does not require any kind of chance or probability, and then summarise it into a proportion which they call

probability*. The idea is that you inhabit one of those worlds but you don't know which.

Two problems arise. First, implicitly, each world is equally likely. This just passes the buck on probability onto the definition of possible-worlds, because no philosopher in metaphysics counts the worlds in a literal way as frequentists do. Second, we *do* know which world we inhabit. For the frequentist, it is impossible to appeal to epistemic uncertainty about worlds that we might inhabit. (ℓ, u) is fixed but the relative location of theta on the (ℓ, u) scale is not known.

It seems strange that, having defined probability as a frequency over possible worlds, the frequentist then defines confidence intervals in terms of frequencies, but forbids talk of probability in that context. It's perhaps an artefact of the CI proponents having been different people to the sampling distribution proponents. Surely we can say that we have used a procedure that has a 95% chance of covering the true value. We might even say that *this* CI has a 95% chance of being one of those happy CIs that turn out to be right.

Suppose then that sampling, which is the only thing that is supposed to change between possible worlds, is deterministic, because no probability can be permitted within the definition of probability. This is possible if you accept that all events can ultimately be perfectly determined. Some might reject the frequentist definition of probability at this point, on the basis of quantum stochastics, which are irreducible to determinism. However, let's proceed and see where it leads if we do not take the quantum rejection.

What makes the samples different in different worlds? Presumably, some minute variation in initial conditions, as in chaos theory. Frequentist possible-worlds can vary in the lower-dimensional complex/chaotic way but not in the higher-dimensional stochastic way.

How do we know what aspects are relevant to hold constant and which to allow to change? It is defined in the study design and protocol, but not always perfectly, and there will be fuzzy edges where different rational observers disagree about relevance. If I "roll the dice" in one replication by holding them, sixes upward, between my fingertips a few millimetres above the table, and then let go, I will have breached the rules, and obtained a draw from a different distribution. But this seems to beg the question again. The set of acceptable replications (irrelevant variations in initial conditions) is a reference class. Anyway, set this aside and let's give the frequentist definition the best chance.

*This is why, historically if not much now, some frequentists were wary of the confidence interval, a calculation which shifted the focus from $\hat{\theta}$ to the co-location of θ and $\hat{\theta}$. It would be a strange statistician who refused any inference via mathematical extrapolation from one study but instead insisted on literal replication, but as we've seen recently from replication work, it can be illuminating and surprising. I too would like to see the distributions of literal replications, but we must make do with a mathematical model.

Where do the values of the initial conditions come from? We cannot give them a probability distribution, and they cannot be drawn at chance. We might be able to talk about randomness, in the sense that there are deterministic processes going on all the time that are irrelevant to our study design. But what is randomness without probability: just values that fluctuate enough — pseudorandom numbers? And if they fluctuate enough so that the next study participant could be any member of the sampling frame, without any preference for one over any other, is that not probability? If we have a set, and we draw a subset from it, how can there not be probability built into the very operation of making the subset? I feel like I have reached the end of the road with the frequentist definition of probability, following the signs to deterministic/chaotic process all the way, and even here, I find probability.

I think that the frequentist interpretation is best put this way: probability is a frequency arising from repeated encounters between a world and some process which returns a numeric value. In inference, it is a frequency over possible worlds with true unknown θ, where the process returns a summary statistic ($\hat{\theta}$ for example). The frequency is everywhere in data(statistic)-generating processes and we can't define it in isolation. In practice, we impose some mathematical model for the frequency of different worlds' data/statistics, which is close enough to be useful for our purposes, a model that is consistent, or at least not inconsistent, with our beliefs.

Now, for the Bayesian, we can imagine possible-worlds too, though it is not essential. In Bayesian possible-worlds, there are various values of θ and $\hat{\theta}$, and when they encounter the statistic-generating process, they sometimes produce the same statistic ($\hat{\theta} = \hat{\theta}_0$). We live in one of these worlds with $\hat{\theta} = \hat{\theta}_0$ and we *genuinely* don't know which one. There is no inconsistency in describing epistemic uncertainty in this setting. In practice, we do not visit alternative universes but rather impose some mathematical model for the various frequencies of different worlds' thetas, which is close enough to be useful for our purposes, a model that is consistent, or at least not inconsistent, with our beliefs.

For the frequentist, there is θ and we say nothing about its distribution; picture it as uniform for convenience. There is a conditional probability for $\hat{\theta}$, the sampling distribution. The interval $(\theta - \phi, \theta + \phi)$ contains 95% of the conditional probability, and by symmetry we can centre it on $\hat{\theta}$ and 95% of worlds will have a $\hat{\theta}$ such that $\hat{\theta} - \phi \leq \theta \leq \hat{\theta} + \phi$.

For the Bayesian, there is a joint distribution of $(\theta, \hat{\theta}$. Marginally, theta need not be uniform but I am only interested in the common ground between the two approaches when they share the estimand of the DGP parameters, so $P(\theta)$ should be "uninformative" (low supremum of Jacobian). We obtain an observed $\hat{\theta}_0$ and then the conditional distribution $P(\theta|\hat{\theta} = \hat{\theta}_0$. It is well known that as the prior, and any frequentist tweaks like penalised likelihood, becomes negligible, they converge.

Note that, for both of these, a world's membership of the true or false CI/CrI depends only on its distance from the other world (integrals of the joint distribution), with no attribute changing, which is resemblance nominalism.

Note also that both rely on the definition of a joint $(\theta, \hat{\theta})$ space even if one approach declines to define any manifold on it (we could think of the frequentist θ dimension as producing a series of unrelated disjoint hyperplanes for different values of θ).

They are true in the same worlds but not, it would seem, in the same way, except that a joint distribution obtained by multiplying $P(\theta)$ and $P(\hat{\theta}|\theta)$ can only be symmetric, so the labels can be swapped without any visible effect, even in the infinitesimal slices of frequentism, and it seems that the ways are trivially different. I conclude they are about the same subject but use different mathematical representations.

Let's close with reflection on different mathematical representations of rotations: axis-angle, Householder reflections, complex numbers/quaternions/octonions, special orthogonal matrices, and finally their Lie algebra(s). It's not too hard to imagine some old argument involving the axis-angle enthusiasts who hold to the realist interpretation of the axis, or that an axis could exist, while rejecting talk of imaginary numbers as fanciful smoke-and-mirrors. Would it not be the right thing to do to put an end to such an ill-informed battle and allow mathematicians to use whatever tool was most useful in a particular context?

8 Personal responses to Mayo's closing questions

"Statistical Inference as Severe Testing" is in large part a defence of well-conducted frequentism.[29] Mayo replaces the long-run, wisely, with error statistics, the chance of being wrong in various ways, which is what really matters to most audience members and can be inflated surprisingly easily; as Terry Pratchett wrote, one in a million chances happen nine times out of ten.

This leads to the idea of severity: given compatibility between a claim (for example, parameter values), a model and data, a procedure should have a low probability of making a claim if there are [Yablo] ways in which it might be wrong, and a high probability if there are not. Severity is like the additional critical layer over and above cookbook statistics, where you, the reader, consider whether it was well-conducted and can be trusted. Mayo to some extent, as the name of the book suggests, unifies statistical inference as severe ... well, something, maybe not exactly testing though. Not everything we do is testing, though Mayo's programme of work is all about tests, decisions and errors.

Now I will consider the closing nine points of her book. Each is a com-

mon criticism of frequentism / error statistical practice, and Mayo offers an intelligent and extremely well-informed riposte to each before inviting the reader to develop their own responses. As a practising eclecticist, I am not the typical reader, but I too found it a useful touchstone. So here goes.

8.1 NHST (null hypothesis significance testing) licenses abuses

No. Nothing licenses abuses. But a culture has developed where people feel they ought to have a p-value in order to demonstrate that they are scientists. Poor education (see below) promotes a cookbook approach that allows PTB* analysts to poke around until they find something exciting. In fact, it is easy to find online data science learning material and blog posts which explicitly encourage this.

The cookbook teaching developed out of a desire to make statistics, the heart of quantitative inquiry, accessible to researchers of all backgrounds, where previously it had been found only in mathematics departments. As Box described first-hand in his autobiography, as statisticians were brought into mathematics departments, they moved rapidly to demonstrate (performatively) their subject's virtue, by increased theoretical content, more theorem-proof-lemma presentation, and less applied content.[33] So, theoretical statisticians were no longer the place for, say archaeologists, to go to learn the craft. Instead, subject-specific applied educators sprang up and sought to make statistics accessible by removing the maths. Unfortunately, the critical thinking and grounding in exogenous theory and experience went out with the bath water.[34]

8.2 Inference should obey the Likelihood Principle

The likelihood principle says that the likelihood $L(\theta|\boldsymbol{X}) = P(\boldsymbol{X}|\theta)$ contains all the information that an experiment or data collection has to say about θ. I would rather say "about the DGP", and then of course the informative posterior tells us about that other estimand, belief about the DGP. But set that aside for now, I'll describe likelihood and the rest follows *mutatis mutandis*.

Mayo and I are on very similar ground here despite initial appearances, because we are both concerned, albeit for somewhat different reasons, with the performance of the analysis at hand, not what might be if it were repeated indefinitely or if we "made a habit of it" (which Mayo amusingly calls "subjunctivist" statistics). I realise that it seems ironic that I, having started out calling myself Bayesian, had to go roaming through a lot of ontology, postmodernism and goodness knows what, with the intention of shoring up my views, only to get to a point shared with the arch-theorist of error-statistical inference.

*Push The Button

The likelihood is specific to a model, so the exploration of model space as well as parameter space should be a concern of the real-world statistician[†]. Nobody in the real world talks about the LP; we've had information criteria since the 1970s and regularization since the 90s. So, no, there is more to quantitative inquiry than a likelihood function (a.k.a. probabilistic model), though it is our best workhorse ("al was haar rug 'n bitji [*sic*] hol"[‡]).

8.3 Fisher and N-P form an inconsistent hybrid

I think we should be clear which quadrant in the table we are aiming to occupy, and choose the method accordingly. We should not mix them together or we will cause great confusion. But I agree with Mayo that Fisher-style p-value reporting can also use power, and decision-theoretic binary N-P tests can also report the p-value. It's just that it raises the chance of confusion.

8.4 p-values overstate evidence against the null hypothesis

This is essentially the same as seeing rejection of H0 : $\theta = 0$ as a straw man hypothesis inside a straw man argument, like a creepy Russian doll. I prefer to deal with reality; I thought that was the point of statistics anyway. If H1 v. H2 is more relevant to their needs, do it. The computer is there to serve you, not the other way around. If the audience would be best served by the posterior probability of θ exceeding some pre-specified threshold, give them that instead.

8.5 Inference should be comparative

But why? I do think that comparisons of H1 : $\theta = a$ v. H2 : $\theta = b$ should be done more often, because sometimes that's what the audience need, yet many analysts force them to stitch together information from multiple NHSTs: H1 v. H0 and H2 v. H0. Let's be helpful.

8.6 Accounts should test model assumptions

Yes, and there are some good methods available, but it can't *in general* be incorporated in the same calculation as the likelihood (because of infinite-dimensional non-parametrics &c. &c.). And, testing *a posteriori* over model

[†]If you reply that $\{\boldsymbol{\Theta}\}|(model = 1) \not\equiv \{\boldsymbol{\Theta}\}|(model = 2)$ and therefore the LP still cannot be violated, then, with respect, you are not a real-world statistician. It is our job to guide people to reliable and useful insights as best we can, which requires the consideration of competing models and sets of parameters (subspaces and manifolds of parameter space, possibly). But see Section 8.6.

[‡]That horse's back is not as strong as we might like, when nasty problems spook it. From Radio Kalahari Orkes, who got it from Dawid de Lange, who got it from FW Reitz's *Klass Geswind en Syn Perd*, generally regarded as the first Afrikaans poem, itself a retelling of Tam O'Shanter.

space, in the severe sense, is not going to be achieved by hypothesis tests, comparativist or otherwise. Instead, we have to sit back and think about the models, and — gasp! — leave the office and talk to other people, those who were at or near the data collection for our inputs, and those who must make decisions based on our outputs.

This is why I especially favour those latent variable models and SEMs in Section 5. We have no guarantees about getting it right, long-run or otherwise, but, for now anyway, this is the art of statistics, and the art is long: we should not be so proud as to imagine that no more useful integrations of parameter value assessment and model assessment will come along in the future.[§]

8.7 Inference should report effect sizes

If that's what best serves the audience as the goal of the inference ... you get the point by now.

8.8 Inference should provide posterior probabilities

Likewise.

8.9 Severe testing is not all you do in inquiry

Testing (as in hypotheses) is not all we do. But I insist on severity, and that requires a more loosely defined kind of test, the bending over backward to prove oneself wrong. Justified eclecticism is not anything-goes.

8.10 In conclusion

I think that my standpoint, combined with the practical experience of working as a statistician, struggling to agree goals, define questions and communicate findings, solves these problems, with no need to throw out frequentism on some of the objections that Mayo lists, or those that I made earlier in this paper. Rather, we can refocus on what is helpful, informative and serves the audience. However, the foundations required turn out to be very extensive, and on some of these topics, I have settled on rare views. I do not expect many statisticians to agree with me, nor do I ask them to, but I do encourage them to think about foundations.

Philosophers of science are not really the audience I have in mind here — in fact, I don't have an audience in mind, this document is more like an audit

[§]You might also object that we can do ensembles of models, Bayesian stacking and so forth. Those are nice, but still contingent on model choice and start to obscure the interpretation into the so-called black box.[17] Any numpty, or computer, can do that; we are being paid to guide data-driven decision-making. Step up to the mark and be a professional, or prepare to be automated out of a job.

trail — and they may well object to various aspects, possibly performatively, for the sake of objecting. Fair enough, we've all got to pay the bills. Also, I don't mind telling armchair epidemiologists and dilettante quants to get back in their lane, so maybe it's my turn. (But then, quality-adjusted life-years are at stake with biostats ...) I am an amateur philosopher, so I have written as good an account as I can for myself.

> *"An amateur is someone who doesn't ram his ideas down your throat"*[35] *

9 Practicalities: teaching

To a great extent, the problems with contemporary quantitative scientific practice that I mention in passing above stem from education. It is no coincidence that I opened this paper with the traditional Snedecor-based statistics course. As a trainer and coach, I have to consider how these philosophical standpoints might influence and guide my teaching.

Simulation-based inference is the focus of a major reform movement in statistics teaching.[36] I have included it in my teaching practice ever since learning about it at the ICOTS-9 conference in 2014. I still have plenty more to experiment with and learn from, but it has been extremely useful to students.

Can simulation unify learning, deep understanding and practice across a justifiably eclectic range of methods? Among the "identify then evaluate" methods, hypothesis tests use "shortcut formulas" (a deliberately loaded name that I borrowed from the Locks[37]) that capture the asymptotic performance of simulating data X_{sim} under $H0 : \theta = \theta_0$, calculating a test statistic $T(X_{sim})$ each time, and counting how many exceed the observed $T(X)$.

Among "evaluate then identify" methods, MCMC and other sampler algorithms do something similar by allowing θ to move around and obtaining a sample from $P(\theta|X)$. Likelihood-based inference does the same without the priors (or with flat priors, if you prefer). Maximum likelihood estimation (MLE) does the same, but restricting simulations of θ to 'heading uphill' to $\hat{\theta} = \sup_{\theta} P(X|\theta)$. Typically, MLE then obtains standard errors for each parameter θ_j by assuming asymptotic normality, where $SE(\hat{\theta_j}) \approx \sqrt{-1 \Big/ \left| \frac{\partial^2 \ln P(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_j^2} \right|_{\theta_j = \hat{\theta}_j \forall j}}$, obtained analytically or numerically. Numerical differentiation means evaluating $\ln P(X|\theta)$ a little way from $\hat{\theta}$ in each di-

[*] ... and I can't resist also sharing this bumper sticker, widely sold in Belfast: "The next time someone calls you an amateur, remember: amateurs built the ark, professionals built the Titanic."

rection to work out the (asymptotically quadratic) curvature, which is also another sort of simulation.

I would like to think that there is a wider future for theory-constrained, realistically complex models, justified eclecticism and simulation-based inference. In professional practice, this is already the direction of travel for some, and I am optimistic. But in teaching, my experience, and that of colleagues, is that students who are not (yet!) intrinsically motivated smell a rat. They know that they must acquire some skill in their own list of must-have methods. For those who would be psychologists, that is factor analysis, Cronbach's alpha, and so on. For those who would be doctors, that is DerSimonian-Laird meta-analyses, t-tests and logistic regression. For those who would be data scientists, that is extreme gradient boosted trees and convolutional neural networks. They know that the simulation they are being shown is a pedagogical trick, a nice-to-have rather than a must-have and that they can simply ride it out without serious consequences and wait for the lucrative bits — the performance. ("Ah, but we respect the old ways, and we disregard his words."*) I'm not sure I can blame them, but I can keep offering an alternative, deeper approach. As for students, so also for non-statistical collaborators and audience members.

10 Practicalities: professional practice

So far, I have described how I took an unusually large portion of time and energy to explore and establish a firm idea of what it is that I do, and how I want to do it. This, I suspect, looks a lot like doing nothing to those of us who work in more mainstream settings where performance (as in theatre, and as in hitting targets) is valued. "Doing nothing" is also Odell's bestselling message.[38] She describes learning how to define boundaries of her work as an academic and artist; there are more parallels with statistics than might be apparent at first. I will close with this because, although it is not so much a corollary of the philosophy of science hitherto as a personal preference, it completes the detail of what I do.

Instead of performing, I work as a freelancer, an outsider teacher*. I can focus instead on close and careful questioning of the client, to find what they and their audience really need, match that to the statistical tools at my disposal, and suggest other approaches. As Monteiro pointed out,[9] the professional's role is then that of the dentist: we must advise what is best for the client, not what they want to hear. They do not want us to be their friends.

*From a Davey Dodds song called, coincidentally, Magpie, first recorded by Red Jasper, then the Unthanks

*In the sense of an outsider artist; I owe this very apt term to my friend and neighbour Matthew Grover.

Of course, as with any kind of consultant, the flashy con artist will always make more money and get more likes. But that doesn't concern us any more. We will be valued and grounded in a group of clients and comrades, which matters more, just as we put family before chasing leads far away.

When teaching over longer periods, or (co-active) coaching, this same kind of questioning can take place. When teaching in a short time, such as a one-day course, I have instead to demonstrate what it is to work this way and hope to inspire. Statistics teaching rarely involves this kind of demonstration, with only one textbook example that I know of;[39] I am convinced it can be extended further, across individual events into a kind of professing without ceasing[†] that participants can drop into. How to make that pay the bills is another question.

This kind of detailed, discursive, non-performative work is not always welcome, but it is an act of caring or "maintenance work". I work locally, in a place (the chalk hill country of mid-Hampshire), centred around interactions with people (not via social media, which is performative in the extreme). These are all parts of Odell's recommendations. As Augé (and others) showed, a *place* is the culture, including terms of interaction, established by those people in that physical location.[7] There are multiple places here, for multiple networks of intersectional individuals, so why not kindly build a *place* for statistical inquiry?

11 Future directions

As I wrote at the beginning, this is necessarily a snapshot of my views in time. Here are some areas that I *know* I am going to be looking into further:

- historic arguments of the ontology of probability, especially any that might have appealed to possible-worlds
- Carnap's logical probability and induction
- identity over time and the connection to aboutness
- clarifying my use of terms like things, objects, meta-stable patterns, simple things and networks of simple things

There may also be other new inspirations and changes that surprise me; I certainly hope so.

Occasional re-writes to condense and clarify the whole text are recorded with a major version number at the beginning of the document (e.g. from 1.4.2 to 2.0). Substantive additions are recorded with a minor version number (e.g. from 1.4.2 to 1.5). Essential corrections and clarifications that cannot wait for the next minor or major version get a minimus version number (e.g. from 1.4.2 to 1.4.3).

[†]with a tip of the hat to *Franny and Zooey*

References

[1] Snedecor GW. *Statistical Methods: applied to experiments in agriculture and biology*, 4th edition, Iowa State College Press (1946): p. x.

[2] Godfrey-Smith P. *Theory and Reality: an introduction to the philosophy of science.*, University of Chicago Press (2003).

[3] Rubin, DB. Multiple imputations in sample surveys — a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association* (1978); 1: 20–34.

[4] Casella G, Berger RL. *Statistical Inference*, 2nd international student edition, Duxbury (2002).

[5] Hájek A. *Interpretations of Probability*, Stanford Dictionary of Philosophy (2019). https://plato.stanford.edu/entries/probability-interpret/

[6] Lyotard, JF. *The Postmodern Condition: a report on knowledge*, translated by Bennington G, Massumi B. Manchester University Press (1984).

[7] Augé M. *Non-Places: a introduction to supermodernity* (2nd ed.), Verso (2008).

[8] Grant RL, Hood R. Complex systems, explanation and policy: implications of the crisis of replication for public health research. *Critical Public Health* (2017); 27(5): 525–32.

[9] Monteiro M. *Ruined By Design*, Mule (2019).

[10] Efron B. Why isn't everyone a Bayesian? *The American Statistician* (1986); 40(1): 1–5.

[11] O'Hagan A, Oakley J. SHELF: the Sheffield Elicitation Framework. http://www.tonyohagan.co.uk/shelf/

[12] Glymour C. Instrumental Probability. *The Monist* (2001); 84(2): 284–301. https://www.cmu.edu/dietrich/philosophy/docs/ glymour/glymour2001.pdf

[13] Desrosières A. *The politics of large numbers: a history of statistical reasoning*, chapter 1, Harvard University Press (1998).

[14] Lipton, P. *Inference to the best explanation* (2nd ed.), Routledge (2004).

[15] Platt, J. Strong inference. *Science* (1964); 146: 347–353.

[16] Centre for Multilevel Modelling. *REALCOM: Developing multilevel models for REAListically COMplex social science data.* University of Bristol. http://www.bristol.ac.uk/cmm/software/realcom/

[17] Gelman A, Vehtari A, Simpson D, *et al. Bayesian Workflow* http://www.stat.columbia.edu/∼gelman/research/unpublished /Bayesian_Workflow_article.pdf .

[18] Spiegelhalter DJ, quoted by Llewellyn Smith, J. An expert's guide to risk. *The Telegraph*, 5 June 2013.

[19] Mayo D. Does Statistics Have an Ontology, Does it Need One? *Error Statistics Philosophy*, 14 April 2013. https://errorstatistics.com/2013/04/14/does-statistics-have-an-ontology-does-it-need-one-draft-1/

[20] Gandenberger G. Gandenberger on Ontology and Methodology May 4 Conference, Virginia Tech. *Error Statistics Philosophy*, 18 May 2013. https://errorstatistics.com/2013/05/18/gandenberger-on-ontology- and-methodology-may-4-conference-virginia-tech/

[21] Glymour C. To Save the Noumena. *The Journal of Philosophy* (1976); 73(18): 635–7.

[22] Lazer D, Kennedy R, King G, *et al.* The Parable of Google Flu: Traps in Big Data. *Science* (2014); 343(6176): 1203–5. https://www.davidlazer.com/sites/default/files/The%20Parable%20of%20Google%20Flu%20(WP-Final).pdf

[23] Nuffield Department of Public Health, University of Oxford. Evaluation of the UK Sugar Drink Industry Levy. https://www.ndph.ox.ac.uk/food-ncd/archive/research-projects/evaluation-of-the-uk-sugar-drink-industry-levy

[24] Emery N, Markosian N, Sullivan M. *Time*, Stanford Dictionary of Philosophy (2020). https://plato.stanford.edu/entries/time/

[25] Horgan T, Potrč M. Blobjectivism and Indirect Correspondence. *Facta Philosophica* (2000); 2: 249–70.

[26] Goff P. *Consciousness and Fundamental Reality*, Oxford University Press (2017).

[27] Rodriguez-Pereyra G. *Nominalism in Metaphysics*, Stanford Dictionary of Philosophy (2015). https://plato.stanford.edu/entries/nominalism-metaphysics/

[28] Gelman A, Hennig C. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society, Series A* (2017); 180(4): 967–1033.

[29] Mayo D. *Statistical Inference as Severe Testing*, Cambridge University Press (2018).

[30] Yablo S. *Aboutness*, Princeton University Press (2014).

[31] Manski C. *Public Policy in an Uncertain World*, Harvard University Press (2013).

[32] Colchero M, Popkin B, Rivera J, *et al.* Beverage purchases from stores in Mexico under the excise tax on sugar sweetened beverages: Observational study. *British Medical Journal* (2016); 352: h6704.

[33] Box, G. *An Accidental Statistician*, Wiley (2013).

[34] McElreath R. *Statistical Rethinking*, CRC Press (2016).

[35] Feldman M. *Give My Regards to Eighth Street*, Exact Change Press (2003).

[36] GAISE College Report ASA Revision Committee. *Guidelines for Assessment and Instruction in Statistics Education, College Report, 2016* http://www.amstat.org/education/gaise

[37] Lock RH, Lock PF, Lock Morgan K, *et al. Statistics: unlocking the power of data*, Wiley (2013).

[38] Odell J. *How To Do Nothing*, Melville House (2019).

[39] Cox DR, Snell J. *Applied Statistics: principles and examples*, Chapman & Hall (1981).