A philosophical standpoint for professional statistical practice

Robert L Grant
(writing in a personal capacity)

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

February 22, 2023

Version 2.0.1 — `git commit f53f712`

Abstract

I present a tour of philosophical considerations in my work as a statistician. I attempt to give a justification for my preferences. Most of the ideas stem from a sense of humility, that the work must serve some purpose, that this service is the measure of its success and value, and that any concepts developed along the way, and tools employed or sharpened, are — following Poincaré — there because they are convenient to the task at hand. Looking into, and shoring up, the foundations of my views has been a great source of new ideas and has led me to revise some of what I believe and do; I am sure that this will continue.

Contents

1 Introduction

> *"Oh. You've gone wacky."*
> — A fellow statistician, in conversation with the author at a conference in 2013.

Statisticians sometimes disagree about the right way to turn data into insights. There are methodological debates, of course, but also philosophical ones, the difference being that individual thinkers can and should arrive at different answers to philosophical questions, because there are no global answers. This paper is a statement of the philosophical aspects behind what I try to do as a statistician: an audit trail of why I might have taken some decision. It is opinionated, and I do not commend my ideas to you, the reader. It is just a record of what works for me. I explain why it is not formally published in Section 2.5.

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 their 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. I feel that more of us statisticians should think about these questions and write standpoints like this one.

1.1 Prerequisite knowledge, terminology and assumptions

In writing it, I have assumed that the reader has 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 high confirmation bias.

I try to (very briefly) 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 philosophy. I use references here for clarification and recommended reading, not for academic completeness or back-scratching.

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. I did not formally study philosophy, so to paraphrase Ken Hom, if I can do it, then so can you. However, there was a period of thirteen years between starting to read about philosophy of science and writing the first version of this paper. As a starting point I recommend Godfrey-Smith's excellent

[*] 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

overview of philosophy of science,[2] but the material that I have had to cover to establish the views in this document is unfortunately more wide-ranging than philosophy of science, and gets difficult at points, especially so around the relationship between frequentist probability and ontology. Assuming the reader is a statistician and not a philosopher, they will also have to read some of the references if you want to follow everything fully.

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

2 Rational and professional practice

This first section of the paper is about what I do and don't do — and how I do it — where those decisions have some basis making them defensible. We could describe the focus as ethics and policy, though that seems a bit grand; I don't intend any teleological implications. The final section is similar but contains only matters of personal preference, filling in any remaining gaps.

> *What is the price of experience? Do men buy it for a song? Or wisdom for a dance in the street? No* — William Blake, 1794

Firstly, I ought to make some attempt at defining what the objective of work is, so that I can, at each stage of this section, consider which choices optimise that objective. There is the matter of paying the bills, of course, but in putting many years of effort into becoming a recognised professional, I am trying to add more meaning to work.

I think the added meaning is that the work involves craft: application of some skill that is not so easily found. A natural consideration in this day and age is the extent to which such skills might soon be supplied by AI. I return to this in Section 5.1, but I think that the human and interpersonal skills are the good bits.

What is productivity in this kind of work? It appears not to be money earnt, papers published or other "metrics". It is tied closely to the concept of professionalism, to living the good life in that professional role by doing your job well. That must mean maintenance and caregiving (see also Section 5.1), not whizz-bang. Like many of my generation, I was shown many heroes of science at school, who were depicted as lone Romantic figures battling against the odds to achieve one singular redemptory breakthrough. This is not only a gross misrepresentation of science, it is inherently biased against those in supporting roles (less likely to be White men), it is profoundly modernist-Hegelian, it piles anxiety on the kids, and it smacks of a kind of dystopian indoctrination to be good workers. Professionalism is a route out of this mental trap, but not a sufficient condition.

2.1 Systems

Since starting my number-crunching career with clinical audit (health service performance indicators), I have been focussed on whole systems rather than single cogs. As an academic, I had some involvement with clinical trials, but the great majority of my collaborations were epidemiological, educational, or health (and social) care services research. I continued to make some contribution to the clinical audit scene by serving on the committee that handed out taxpayers' money for such project, and pushed the providers to up their methodological and communication games.

To get anywhere down the narrow path that leadeth unto understanding in these systems, we need a plurality of perspectives. We could, of course, reduce the system to a simple measure, and indeed for statistical analysis we must do some of that, but we must also not lose sight of the fact that we are looking only at one limited aspect.

It is heartening to see that "systems thinking" has become a hot topic in the public sector in recent years, although it is too early yet to know if it is just the latest must-have training trend without any sustained impact. My experience taught me that that we need to be clear — very clear — about what we are trying to achieve before we get started. Analysis is always for some purpose, some decision-making, some stimulation of change. It does not sit alone as most people in a scientific education are taught; it is not a virtue in itself to file away some obscure numbers in the hope that a future scholar will think of something useful to do with them.

Qualitative data and analysis has an important role to play in understanding the system, as Rick Hood and I wrote in a paper on public health research and replication.[3] The qual informs the quant and vice versa. This is often the most effective way of widening the set of voices and perspectives that contribute to the research, though it is slow and effortful, and skilled qualitative researchers are in short supply.

2.2 Humility; serve the audience

In teaching almost any subject in my field, I usually start by showing data analysis as the middle part of a three-part process. Our data are usually not collected or designed by us, and sometimes are a devil-may-care collage of databases and sources that happened to be lying around. If we want to do good analysis, we must understand the source, and the best way to do this is to leave the office and talk to the people who did the collection. They often have eyebrow-raising stories to tell about shortcomings in the source data. (There is little point in asking their boss — the boss is always told that everything is fine and dandy.) By building knowledge of those shortcomings into the analysis, we can get more reliable results that do not flop or end up embarrassing everyone down the line.

The other end of the process is that our outputs go to someone else (the boss, the client, colleagues or the public) who uses them (we hope!) to make better decisions. What do they actually need? And how much numerical detail can they handle? We need to take this into consideration when we design the analysis too. There is no right methodological choice until we have considered these two parts of the process between which we are sandwiched.

This comes as a surprise to anyone schooled in the common, mechanistic, flowchart, push-the-button approach to statistics, which is increasingly criticised nowadays, most eloquently by Richard McElreath.[4]

In one of his rather good videos, Rasmus Bååth gives an example of a fishing company that must make a binary decision on whether or not to send free samples when they launch in a new market or not.[5] If they do, there will be fixed costs, so what the boss really needs to know is:

> What is the probability that we will make back the cost of the free samples via higher profits?

That would require a posterior probability of the predicted profit exceeding some threshold. It would not be helped by a p-value testing whether or not the true added profit of the free sample is zero. It would not even be helped by a confidence interval that tells you one of these:

FREQUENTIST: Here is a 95% confidence interval. If we were to repeat the data collection process many times, we would obtain many 95% confidence intervals, and 95% of them (asymptotically) would contain the true additional profit.

BAYESIAN: There is a 95% chance that the true value is in this interval.

(More on this frequentist and Bayes stuff later. So much more, in fact, that you may lose interest in the whole thing.)

The posterior probability would be simple enough that even Dilbert's boss could handle it, so it works in terms of communication too.

The trouble is that the boss (or whoever) typically does not have even this level of clarity. We need to help them to get there, and that is something I always devote serious chunk of time to. They sometimes find it frustrating, but I am there to be a professional (I'll define that in Section 2.3).

Of course, we need to define the question even further. We need to know whether it will make back the outlay *over what timescale* and the threshold might not be the precise outlay but something else that the boss needs to be able to show off at the board meeting. We might not know about that level of detail, and they might not have thought of it yet, but we can get there together.

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 3.7.

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.[6] 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.[7]

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 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*.[8]) 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.

*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.

2.3 Professionalism; communication; confidentiality

To be a professional statistician is important to my practice. I am a chartered fellow of the Royal Statistical Society, which gives me that air of gravitas to defend my advice to a client. I think it is important to have such societies, and for statisticians to support them. They are not synonymous with academic societies; they are regulated in some way by government and set high standards for membership, thus acting as quality assurance. It is a pity that people entering data analysis via machine learning or data engineering do not yet have such structures to support them; I hope they get there soon.

In my mind, I associate my professional role with Mike Monteiro's analogy of the dentist.[9] When I first encountered one of his talks online, I mentally substituted designer for statistician and found that it all related to us too. When you go to the dentist, you want your teeth sorted out. You are not looking forward to it, and you hope there will be nothing serious, but you recognise that it needs to be done. That's how the boss feels about us.

The dentist is not there to tell you what you want to hear, but what you need to hear. You might hate it at the time, but you'll be grateful later. The statistician should be able to confidentially talk to the boss about analytical projects that are doomed or misunderstandings that need to be discretely cleared up before things go wrong. I am not interested in work that does not have this professional relationship.

It is not the same as producing glossy charts to back up what the boss wanted to do anyway, nor coding up some random forests in Python, not for production but just so that the boss can tell their buddies at the golf course that they are leveraging AI.

You may have seen one of those ever-popular data science Venn diagrams, which define what a great data science team should include in its collective skill set. (Some, erroneously, attribute the full intersection to the perfect data scientist, but this is foolish: anyone who had spread their learning and experience that thinly would not be very reliable.) Communication skills are always there in the mix, and as a consultant and trainer, I hear frequently from organisations that they wish that they could find analysts with communication skills.

I feel that the professional statistician must take an interest in communication of their findings: how to verbalise, how to visualise, how to explain methods for non-technical and time-pressed audiences. It is a never-ending mission to refine these techniques, but it is a satisfying one. No matter how clever one's calculation is, if it is not understood, recalled and acted on, then it was a complete waste of time. Crucially, if you don't do it, then one of two things will happen. If you are lucky, nobody will notice your work and they will just keep paying you, until the next recession anyway. If you are unlucky, someone else will explain it for you, and quite likely make you look stupid in the process.

Another vital aspect of professionalism is confidentiality. We have to take this very seriously. There's plenty that I do and never talk about. Of course, it comes hand in hand with the responsibility to weigh up the ethics of whether to deal with the client at all.

Confidentiality is also a responsibility toward the subjects of the data. I don't think this ends with the letter of the law. I take a strong definition of consent: actively opting in is the only way to be legitimately analysed. Here there needs to be a balance in settings of common good, which potentially covers a lot of healthcare statistics, but I have always consistently seen researchers, bureaucrats and healthcare professionals trying very hard to evade the regulations in this respect, to jump through hoops and say the right things, and then to do whatever they like, or to get some kind of exemption.

I also take an unusually strong stance on human remains and tissue. UK law says that people cease to be "natural persons" when they die, and then you can do pretty much what you like. That seems profoundly at odds with my culture*. Where there was no proper consent, I think we must leave it alone, and I should refuse to participate. I used to teach with the Titanic passenger dataset, as many do, until I thought this through. Then I decided only to use the version without names, and preferably not at all. Actually, I haven't used it at all since then. I found on a social media platform that several other stats teachers felt the same way.

I teach people on my data visualisation courses that they will need to do three things that are uncomfortable when you come from a scientific training: curation, compromise and consultation. By consultation, I mean you go and ask other people what they need and where the data came from — we already covered this.

Curation, because I can't communicate everything most of the time, and have to decide what to highlight and what to leave in the lengthy methodological appendix. This is always awkward, but it has to be done. The professional must deliver this service.

Compromise, because I sometimes know that my audience will not fathom or accept what I really want to show, so I meet them halfway in a responsible way. This might do me something of a disservice, but it should not do them a disservice. To dig my heels in, or to give in to their whims, would be to do a disservice. I am often educating alongside reporting stats.

For example, in one paper where the (ill-advised) standard way of analysing the outcome scale gave Mann-Whitney p-values, and I then did a Bayesian structural equation model, my collaborators wanted a p-value for comparison.[10] I could see that refusing might have led to the SEM disappearing into

*Perhaps I feel more linked to this ancestrally as a Scot with a passing interest in the Stone Age. I would have excavated ancient remains put back too. So what, you feel curious? Grow up.

an online appendix or "second paper" of the sort that never actually gets done. So, I noticed that the marginal posterior distributions of the main effect statistic were nearly normal, and not far from zero, and that the priors were $\mathrm{N}(0, \sqrt{1000})$, and concluded that, had we done some magically equivalent maximum likelihood analysis, we would obtain p-values which were the tail density of that same distribution beyond zero ($\times 2$). That's a fudge, but one that does no harm to the collaborators or the audience. There's more discussion in my PhD (by publication) introduction.

2.4 Strong inference and inference to the best explanation

The task of statistics is always a little open-ended. Yes, the client, boss or whoever has some question and we serve them in that regard, but as we provide results for one question, others open up. Sometimes we are testing hypotheses and sometimes we are generating them. Some statisticians look down their noses on the generating part, but I think this is unwise. Like with communication, if we don't do it (or at least contribute to it), someone else will.

I like Platt's notion of strong inference,[11] where scientific learning is an iterative process of induction and deduction, repeated as we zoom in on (but perhaps never quite reach) insight. The reductive tradition of running a Neyman-Pearson-Wald test and declaring the findings significant or not and then calling it a day is not only depressingly pedestrian and unchallenging, it also short-changes the client. Shouldn't you help them with what might come next, even if they are not quite ready for it yet?

Note that I don't object to the test itself, as some people do, and I will return to this in Section 3. I object to the calling-it-a-day.

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*.

*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.[12] 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

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".[13] 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 understanding, perhaps by matching other external results or related phenomena).

This inference to the best explanation is not the same as causal inference; 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.[11] 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.

2.5 Lyotard postmodernism

I mentioned the need for multiple viewpoints and voices in my system-focussed work above. I think this is the only sensible way to maximise understanding. We cannot take the role of the privileged viewer who is above bias. Instead, I want to construct analyses of the data that are suitably complex to accommodate complex phenomena ("realistically complex"[14]). This starts to push me towards Bayesian methods, latent variables and giving emphasis to careful human interpretation of results. It pushes me away from hypothesis tests, black-box machine learning and automated AI implementations.

It also pushes me toward humility and away from performance. 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.[15]

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.[16] His modernism is the period in European / North American culture when a privileged class of people are permitted to practice science by

independent fact in a constellation, the form of which only the Aesclepian initiate may discern.

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 Xhosa, *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, I describe my work as 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. I also try to accept and work with uncertainty, in all senses, including Charles Manski's.[17] I think we should accept that judgement of the reliability of findings — and generation of hypotheses — are subjective and difficult (error-prone), and get on with it anyway.

This led to my paper with Rick Hood on trouble brewing within public health research.[3] 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. We might call them "meta-stable patterns", an idea and a phrase that I will return to later.

To research this sort of system requires, first and foremost, the voices and views of those who are within it. 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.

That's why this document is not published in any formal way: I do not ask anyone to read it (unless they want to unpick why I did things a certain way: the audit trail), let alone follow it, nor do I expect it to stay the same for long.

3 Inference, probability and metaphysics

This section of the paper deals with the old question of what probability is, and hence what kind of statistical inferences and conclusions we can draw from our analytical work. It is often framed as frequentists versus Bayesians. On reflection, I think this is not quite the right classification. My aim is to get some level of philosophical rigour, to give my preferences a run for their money.

In summary, I have used an eclectic approach to inference, applying Neyman-Pearson-Wald tests or maximum likelihood estimation or Bayesian modelling, according to the needs of the task.

There are those who hold the view that only one interpretation of probability is correct, though this is less common as time goes by. Increasingly, people use a bit of this and a bit of that, but being *à la mode* is not a guarantee of good practice. Many do not consider whether there is a philosophical problem in casual eclecticism at all, as long as it works, and some are even proud of that*. I am not so sure; if clever people like Deborah Mayo or Stephen Senn might object to my work (which is in the eclectic camp),[18] then I ought to think carefully about whether to mend my wicked ways, and

*for example, when machine learning people reject consideration of statistical, let alone philosophical, niceties on the grounds that their work — I don't know — makes more money? is in Forbes?

if not, then I should prepare my defence.

First, let's be clear that it is unhelpful and potentially misleading to present results that have to be interpreted in various different ways. For example, in network meta-analyses it is common for pairwise comparisons to be done frequentist-style, and then the combined analysis Bayesian-style. The problem is that such casually eclectic analysis involves mismatched philosophical foundations, so the outputs cannot be contemplated together.

I contend that my approach is *justified eclecticism*. I do not apply a given inferential method (and philosophy of probability) because it works in the circumstances, as the machine learning community often does, but instead because the question that is being asked is best addressed by it, and I apply only one interpretation of probability to a project (although some aspects of projects have at times been beyond my control). The method and implied philosophy is chosen to best serve the data, the audience, and the question. As a result of carefully considering this, I arrived at a new kind of taxonomy of methods. The rest of this section presents the development of these ideas in logical order.

3.1 Different inferential procedures from my perspective

Bayesian statisticians can describe procedures that use only likelihood as a special case of Bayesian inference: one where the prior is flat. (Whether they object to the flatness is another, more practical, matter.) There are also objections from Bayesians to Neyman-Pearson-Wald null hypothesis testing (setting aside the popular practical objection that it is often abused), typically that the calculations are contingent on specific parameter values which are "known to be false". That objection, from Bayes to NPW, is one I will return to later in the broader context of scepticism. It doesn't bother me much as it also seems logically flawed: frequentist testers do not make any claims about knowledge or truth.

Justified eclecticism starts with the view that no one interpretation of probability has a valid claim to exclusivity. If there is a strong claim to exclusivity, to having the one true version of probability, then it is made by frequentism. Commonly, it rejects Bayesianism, and some extreme individuals from time to time have also rejected likelihood-based inference.

Figures 1 to 4 show how I think about these distinctions. I have not seen this kind of exposition before, so I think it is worth setting out here as the basis for the rest of this section. In each graph, we have theta-hat ($\hat{\theta}$) on the x-axis, the estimated or putative values of an unknown value which we seek to infer. The possible true values of that unknown, theta (θ) are on the y-axis.

In the most frequentist of procedures, Neyman-Pearson-Wald testing, the analyst must propose one or more values of θ, and then calculate the probability of obtaining data that leads to the estimate $\hat{\theta}$. Hypothetical

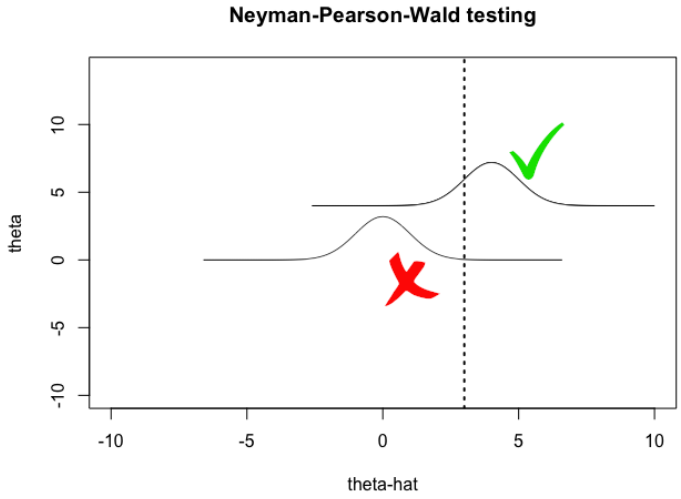

Figure 1: Two hypothetical values of θ (0 and 4.4) are compared using a model for the distribution of data, given this unknown value. The data has provided an estimate at $\hat{\theta} = 3$, which leads to preferring $\theta = 4.4$.

values of θ can then be "rejected" or not.

Often, only one value of θ is proposed (the null hypothesis), and an *a priori* value of the probability $P(\hat{\theta}|\theta)$ is used to reject it (or not).

We might also calculate the area under the curve and so evaluate an interval (perhaps open) of θ. This appears, for example, in one-sided tests, non-inferiority or equivalence clinical trials.

A rarer approach is to specify two points in the domain of θ and find which is better supported (has higher $P(\hat{\theta}|\theta)$). This last case is shown in Figure 1.

Likelihood is, of course, the name we give to $P(\hat{\theta}|\theta)$, because in practice, the data are fixed, along with the estimate of $\hat{\theta}$, while we want to consider alternative possible values of θ. In frequentism, probability only makes sense as a proportion of events in a long run. I will return to the long run below, but for now, we need to be aware that frequentist use of likelihood does not permit θ to be seen as a continuum, but instead as a set of values, which may be infinite but still have gaps between them. Mathematically, the frequentist θ is a metric but not continuous space (the θ in these figures takes any real value, so it could be metric; some unknowns, such as discrete, ordinal latent variables, are topological but not metric; I present this real one because it is easier to visualise and think about as an example)[†].

[†]If you have not read any topology, just ignore any talk of continuous, metric, or topological spaces that might appear in this section; it is helpful, but not essential to understand the rest.

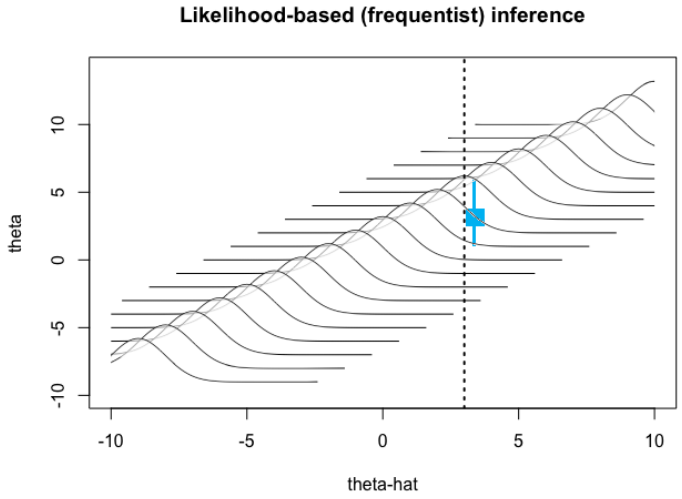

Figure 2: There are an infinite number of possible θs, but not a continuous space of them. $\theta = 3$ obtains the highest likelihood and is the maximum likelihood estimate, while a confidence interval around this can be calculated to satisfy one of the typical criteria, even in strict frequentist terms.

This situation is shown in Figure 2. The number of putative values of θ is tending toward infinity. Nevertheless, its definition remains frequentist: there can be no probability over it, only conditional on it. We can find a set of θ values which satisfy some criterion based on $P(\hat{\theta}|\theta)$, and so add confidence intervals to our repertoire of significance and p-values[‡].

Bayesian statistics, with a flat prior, adds to this by allowing probabilities over θ as well as $\hat{\theta}|\theta$. Note that, using basic probability theory, this means that we are specifying a bivariate distribution over $(\theta, \hat{\theta})$. There is now a continuous metric space over θ ($\subseteq \mathbb{R}$). We see the result in Figure 3: the likelihood is (proportionate to) a horizontal slice through the bivariate distribution, and the posterior is a (normalised) vertical slice through it.

Extending this to a non-flat prior is relatively simple: by adding a distribution on the y-axis for θ, we obtain a proper bivariate distribution (Figure 4); inference is as before mathematically, but interpretation will be quite different depending on what is *meant* by the prior: the definition of probability. One analyst in one project might opt for personal belief, another for consensus belief, another for previously published frequentist evidence (asymptotic

[‡]There were some antediluvian arguments about significance (decision-theoretic inference associated with Neyman, Pearson and Wald) versus p-values (a more iterative, touchy-feely process associated with Fisher). As Deborah Mayo has said,[18] this is a false dichotomy, because there is no philosophical or mathematical barrier to the Waldist using p-values for degrees of certainty or the Fisherite using power calculations and typing asterisks hugger-mugger. The difference is cultural and I will leave it aside.

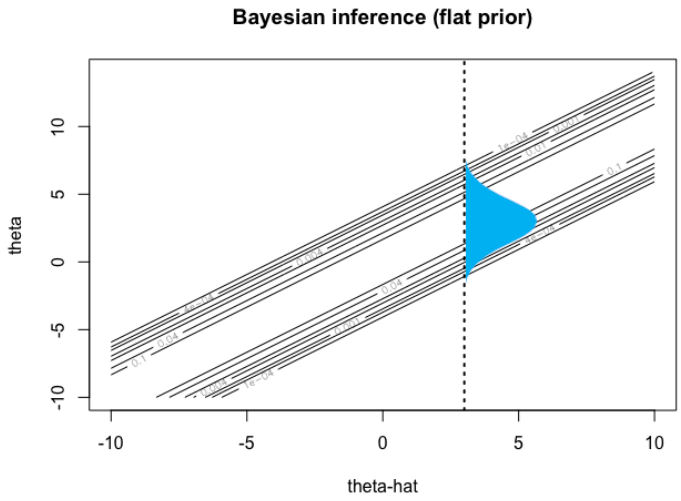

Figure 3: Bayesian inference with a flat prior allows a bivariate (improper) distribution; we see the contours of this here. A vertical slice at the observed $\hat{\theta}$ leads to the posterior. Now, we can talk about the probability that θ has certain values or lies in certain intervals, by integrating this posterior.

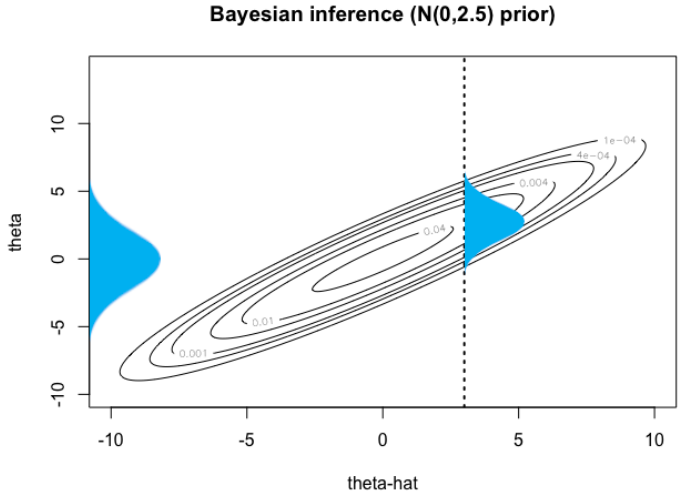

Figure 4: Bayesian inference with a normal prior.

sampling distribution), yet another for a heuristic computational aid that adds no meaning over and above the likelihood.

This highlights for me the fact that the mathematical procedure seems to categorise the methods at one set of cutpoints, while this definition of probability seems to add different ones. It is certainly not as simple as "frequentist versus Bayesian". I return to the implications of this below.

3.2 Strengthening frequentist foundations

Now, if the most likely source of philosophical criticism of my work is exclusive frequentism, then my task is to give that its best shot, and to see if I can withstand it. There are, unfortunately, a number of strange things which are said about frequentism, both by its adherents and detractors, and I would like to try to clear those up first, to strengthen frequentism as best I can. It is a pity that statisticians do not, in the main, think about these foundations very much. We are human, after all, and have to be effortfully on our guard all the time if we are to avoid the cognitive bias that venerates *ipse dixit*.

3.2.1 Fallacies

All alternatives to frequentism extend beyond the long run of events, and so allow for uncertainty arising from sources other than sampling error (which is sometimes called aleatory uncertainty). Anything that is not aleatory is epistemic: it is uncertain simply because you don't know it. Of course, there are Bayesians who counter-argue that *everything* is, at heart, epistemically uncertain, implicitly making some appeal, which they have generally no understanding of — nor, in my experience, inclination to acquire it — to the innate nature of reality (it is conceivable that some of these arguments could be on the basis of religious belief, though I don't recall ever hearing such a claim), but I will leave this aside; we will need firmer arguments than that.

Some people talk about permissible forms of uncertainty in terms 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 would retrospectively become invalid, like in quantum entanglement. This, it seems to me, is not what they 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 θ. 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.

So to the long run. There is a canonical example of election polling. If an event only happens once, and we are estimating what the summary statistic will be before the fact, then, the argument goes, there is no run at all. One might conclude that a strict frequentist should not publish any inference in this setting, such as a "margin of error" (and indeed, some people out there say this). Now, there are certainly occasions when I decline to do any inference, and that is when the data are *census*: the complete population. The only role of statistics then is descriptive, which provides the whole story.

But, sometimes, the question which we are trying to answer is a little different. Problems arise, as ever, when it is not clearly defined for all. Is the election poll estimating the outcome of the election, or the prevalence of different intentions in the population on the day of the poll? Inference to the latter is simple, but to the former requires an additional, interpretive, contextual, non-statistical leap.

The election only happens once, but the election is not the data collection process: that is the poll, which surely can be repeated (and usually is).

Another argument (implied by the election example) appears from time to time about whether the data collection process can be repeated *indefinitely*. Let's think carefully about what it means. A repeated data collection is a second instance of the same process, which leads to another sample of data. Nobody intends it to actually happen (other than vexatious and philosophically ill-informed anti-frequentists): it is a potential event. This clarifies that it is not a literal sequence of data collections in the real world. Any argument that the (presumptively infinite) frequentist long run is a fiction because of the inevitable expansion of the Sun and destruction of the Earth is missing the point, and does not provide any support to Bayes. I, too, used to make this argument but now I know better.

We should focus instead on understanding precisely what is meant by this vague term, the "long run".

3.2.2 Possible-worlds and boundaries of the sampling space

I think that the correct concept from philosophy is that of *possible-worlds*: there is a set of possible data collections that we can imagine, all done in identical worlds but not exactly in the same way. It is not literally identical, or it would happen at the same time, in the same way, and obtain the same sample. Our real-world sample is one of them, and we don't know which one. Our estimate ($\hat{\theta}$) is one draw from the sampling distribution, but we do not know which. Hájek responds to, and summarises the possible-worlds concept

of frequentism,[19] although I do not agree with everything in his critique.

Many statisticians, and consumers of statistics, have never thought very closely about what they mean by the long run. I keep thinking about a commentator after the end of the 1995 Rugby World Cup, who said that "9 times out of 10" South Africa (who won) would have lost that game. It is a funny thing to say, but everyone gets it. That's frequentism all over.

Rather than say "long run" and invite the listener to imagine a sequence in time, it would be more helpful to talk about *sampling space*, which contains possible-worlds. Some worlds are in the region of sampling space that constitutes a relevant sample, others are not, and some will always be contentious, I contend*.

Clearly defining the question in frequentist analysis requires clarity in turn about the long run. We know that there are parameters — unknown values sought by our analyses — which, if we were to repeat our data collection process $n \in \mathbb{N}$ times, and estimate the parameter each time, would always take different estimated values, drawn at random from the aleatory sampling distribution (which arises from the possible-worlds sampling space).

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 unknowns should be impossible. Epistemic unknowns such as missing data are simple in this regard: if we repeated the data collection in another possible world, we would not obtain different values of the same missing value; we would obtain entirely different missing data, or maybe none at all. The only way to infer them is Bayesian.[20]

But what about unknowns in between, a grey area including group-level effects in a multilevel regression? In some circumstances, for $n \leq \nu \in \mathbb{N}$, 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. Why do so unless you really believe there is no probability except for in literally eternal runs? However big ν is, this kind of frequentist wants more. †

*If we are analysing the fairness of flipping a particular coin, say an American quarter dollar, then we would obtain some of these coins and flip them. If you propose a possible-world in which the quarter dollar is a different shape to the real-world, it would clearly be inadmissible. If you flip them in a different way to me, maybe just throwing them at the table lightly, then some rational impartial observers would object and others would not. If you proposed to flip them by holding each one a half inch above the table and dropping it, then everyone would object. It seems to me that a contentious zone can always be concocted, though proof remains on my to-do list.

†This is a kind of virtue signalling for some people. It reminds me of when my wife and I were in a café in Pittenweem, a quaint village latterly gentrified by Edinburgh commuters, and I found a copy of a research report related to the village's arts festival lying around.[21] The author must have struggled to keep a straight face as they typed *We asked how long your family had lived in Pittenweem. The answers ranged from "6 months", through "longer than you" to "hundreds of years".*

The status quo however does allow estimation and inference on related parameters that do not appear and disappear as n increases, including the variance of group-level parameters ("random effects"). (Note that they are "effects", not parameters, so as not to upset sensitive ears). 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 strange semantic contortions!

We need to consider the implications of not following me in accepting the possible-worlds interpretation. In this case, it seems to me that some subjects of study have quite different limits, ν, on aleatory re-sampling, depending on the question being asked and the realistic data collecting process. The fact that the process can impact on inferential procedures was already evident above in the case of census (as in whole-population) data.

If the groups in the multilevel model are learners taking a one-day course, we expect them all to be different in tomorrow's data, so $\nu < 1$. When ecologists count species along transects, and the transects are the groups, we expect the ecosystem in each transect to persist, but not in every possible-world sample; in some, what was a forest will be a desert: $1 < \nu < \infty$. The sampling space is a subspace of the possible-world multiverse.

Transects persist, while learners 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. A large ν would warrant frequentism, a small ν Bayesianism (or nothing).

The fact that this does not happen is, for me, a symptom of the lack of philosophy among statisticians and quantitative researchers. Also, it might be (I'm not 100% sure, it is not a priority for contemplation time) that an exclusive frequentist had better have an acceptable lower value of ν up their sleeve, call it ν^*. Perhaps you only do inference if $\nu \geq \nu^*$. Of course, you don't really do repeated data collections in real life, so it is a question of assessing the probability $\mathrm{P}(\nu \geq \nu^*)$, which should exceed some threshold, like 0.5. It's their call, but they would need to stick to it and not allow variation by even $\nu^* - 1$, or they will fall into a sorites paradox.

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 uncer-

tainty 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.

I want to return to the problem of semantic propitiation: 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, all just so that we don't bring down the wrath of the watchful exclusive frequentist upon us. I don't like this sort of posturing very much, as discussed earlier in the context of performativity and postmodernism.

Interestingly, the concept of significance, though it is typically tied up with p-values, is not the sole preserve of frequentism, but just a dichotomising cutpoint for binary decision making. Under the right circumstances, we expect our error rates to be well understood, so why not pre-specify it in terms of a posterior distribution instead of the p-value?[‡]

There is a half-reasonable objection in terms of severity,[18] that including priors adds another potential banana skin to the process, but that in itself does not even guarantee that a Bayesian model will on average, over all efforts and projects, have inferior error rates to a Neyman-Pearson-Wald test. What matters is how they are done, the human element, which is the source and measure of severity[§].

3.3 Reference classes and the ontology of frequentism

Frequentist analysis, if it has a claim to exclusivity, has it on the basis of the set of possible-worlds samples. This then provides a proportion (numerator over denominator), which is probability. The claim would be that there is a real (at least potentially via repetition of the data collection) numerator and denominator, so probability is real (at least potentially) and does not require fanciful talk of belief or propensities or logic. It is a real thing because the numerator and denominator are, at least, in theory, within our grasp as unequivocal numbers, actual real counts.

This realist frequentism relies on a realist reference class, which is the denominator. For it to work, it must be universally clear whether a particular data collection is in the reference class or not. Obviously, there is some wiggle

[‡]There are several interesting (I think!) experiments along these lines in the papers that make up my PhD (by publication) portfolio. They did not all work, in that they did not all inspire deep understanding in collaborators and audience, but some did, and I'm glad on the whole that I did them. They are fundamentally experiments in communication.

[§]Someone once said to me that when there is a `severeTesting` R package, they will believe it is a real thing. I don't agree but I can see where they are coming from.

room around the usual phrase, "identically repeated experiments", which I think would be better named as nearly identical data collection processes (nidcops for short). If they were identical in all respects, the data would be the same too. But how different is too different? I suggest that the boundaries are contestable and fuzzy in each analysis, which is to say that there is scope for disagreement among rational, well-informed, disinterested analysts.

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 really 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 use Kelvins as a unit of temperature for that reason?

Quite different to this are 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.

For me, this is familiar and, frankly, perfectly acceptable. It is the same contestable interpretation that we have in choosing a model for the data generating process, or defining the question, or generating the next hypothesis. It is hard, and unclear, and all we can do is try our best. It is misleading to suggest that there is a computational flowchart or algorithm that will take us straight to truth, or even truth most of the time.

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.[22,23]

Glymour pointed out that realist and antirealist standpoints on theory and reality both have problems.[24] 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.

This offers an alternative approach. Realist sample space and the long-run might hold for simple, more fundamental, things, like protons, 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. It is hard to pin down the boundary to the sampling space. Horgan and Potrč amusingly call the simple things snobjects, and the fuzzy aggregate concepts slobjects; we will encounter their third class in due course.

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. But that is a quite different thing, mathematically and physically.)

And yet, this Spiegelhalter-style sceptical probability still manifests as an asymptotic frequency over long runs*. I begin to feel that frequentism is a good and useful method, but might not have a claim on exclusivity. First, though, I need to clarify exactly what my ontological stance is, instead of just tearing down everyone else's.

3.4 Resemblance nominalism

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 reference class? You might pick a point, but you cannot guarantee that all rational impartial observers will agree.

And also, there is fuzziness arising from problems of aboutness.[29] The scene is in a museum:

Dad: There are no quaggas any more.

Daughter: Yes there are, like this one. Stuffed in museums, ha ha.

Dad: That's not what I meant.

*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 in parallel universes, perhaps with identically replicated researchers too.

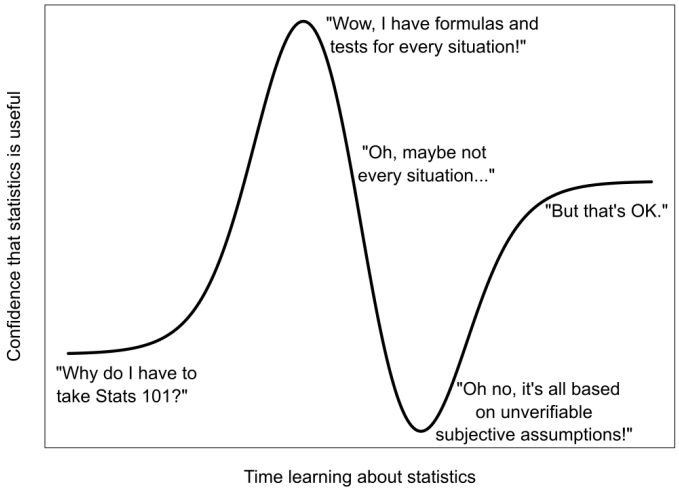

Figure 5: Learning the practice of statistics is full of surprises.

Kids around the age of six have a special power for undermining reference classes. Later, they learn (we learnt) to take language seriously (big mistake, in my opinion) and to believe in the certainty of science, which appears to deliver truth if we just push the right buttons. Figure 5 is my version of the famous Dunning-Kruger curve, applied to statistical education, which I hope will appear in the preface of my forthcoming book with Gian Luca di Tanna, *Bayesian Meta-Analysis: a practical introduction.*

Aboutness reappears in Section 3.6. Could it be that classes, like magpies, quaggas, and indeed nearly identical data collection processes, are just fuzzy semantic conveniences (FSCs)? I'll try to firm this up, and look into all the potential counter-arguments. Remember that it is a philosophical issue, so there will not be one universally proven answer, just one that I have settled on.

If we reject the realist-frequentist 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 probability as an asymptotic proportion 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 3.6.

Clearly, one way to reject realism of frequentist reference classes is to reject all reference classes. This asserts that classes of things do not really exist, but are just a semantic convenience. We humans made them up. In the

philosophical field of metaphysics, this means that you are not a realist but a nominalist. I find this quite natural, through an unusual route described in Section 3.5, but however you find your way to rejecting the realism of the sampling space reference class of possible-worlds, justified eclecticism (Section 3.7) follows.

In metaphysics, a common subdivision of realism is over whether properties 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, and hence are somehow allocated by a kind of perinatural 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 nearly identical data collection processes (nidcops), then there must be a property of belonging to that class, which discriminates the repetitions that are admissible from those that are not. For the class of magpies, there has to be a property of magpiety, so where is it? These properties are kinds of *universals* in metaphysics; the data collections and magpies are *particulars*. *Ante rem* realism is tolerated in metaphysics because philosophers also want to solve puzzles like "where is when?" or "is 7 up?"*.

Whether there are abstract objects or not is not really relevant to my standpoint on statistical practice, 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 and resemblance nominalism 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 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 impartial observers (a pleasing and useful phrase from Tony O'Hagan and Jeremy Oakley, in a different setting[7]) would agree on the numerator and denominator that arise from it. When they don't, it is a FSC†. And I still contend

*One of these is proper ante rem metaphysics, the other is a track title from a 1980s acid house album. Or is it?

†You might complain here that I am letting man be the measure of all things, and so begging the question, but I think it is a more thoroughgoing problem than just the irreproducibility of *decisions* on the admissibility of nearly identical data collection processes. 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

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. The magpie/soil/hawthorn is so riddled with conflicting *in re* properties that overlap and confuse us mere humans that we have no chance that all of us will always understand the full story (that's what would be required for realist reference classes).

Of the nominalist standpoints, I favour resemblance nominalism, which says that these black and white descendants of dinosaurs 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 a possible-worlds reworking of the arguments. 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.

So, I conclude that there is no way of defining the permissible sampling subspace / reference class. This does not mean that I reject the use of frequentism. Far from it! I find it useful for the right job. But I do reject the claim of exclusivity.

3.5 Existence monism

For me, resemblance nominalism follows logically from an ontological stance called *existence monism*. This is an admittedly unusual view (like resemblance nominalism) that claims that, in realist terms, there is only one object, the entire universe. All smaller objects are mental constructions.

It is more helpful to think of the universe as one unfathomably convoluted object in space-time (Horgan and Potrč call it the "blobject"[25]). You may name particulars and universals as you wish: they do not exist except in your mind. Nevertheless, that can be a useful thing to do, and we humans,

---

imagine anyone making a sensible argument that there is never any such scope.

[‡]The choice of the letter F is Goodman's, and is used by other metaphysicists, but I enjoy the suggestion of Fuzziness

[§]*viz* joelcadwell.blogspot.com/2014/03/warning-clusters-may-appear-more_23.html or washingtonpost.com/opinions/2021/07/07/generation-labels-mean-nothing-retire-them/ — but also, as you will by this point expect me to write, because *no* label has any basis in reality.

and many other animals besides, would not have progressed much without these (fast and frugal) mental constructs.

When a baby (or a kitten) learns that a magpie flying behind a tree and emerging again is, in fact, the same bird, they are acquiring one of these helpful constructs. But there is no magpie really. Not *really*. There are cells, and some of them are non-avian bacteria in the gut. And look inside, what about those mitochondria? Is even the cell really one thing? Rational impartial observers might disagree. Over time, it arises out of atoms, arising out of the nascent Sun's accretion disc, arising out of other atoms and subatomic particles inside a long-ago star, arising out of quark-gluon soup, arising out of the post-Big-Bang something-or-other. It is a pattern in a space-time swirl of bits arriving and leaving, just like you and me.

We notice 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). Helpful metaphor: we see and count waves on the ocean, but there are no waves really, only the ocean.

Objects can only be named, 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 data or statistics arising from nearly identical data collection processes are actually just meta-stable patterns.

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 scattering and gathering that is going on.

Reference class names are neither objects, nor properties — because they do not make subclasses of objects (there being only one blobject). They are not intrinsic but a subjectively defined description, an utterance of one part of the blobject about another. The frequency follows directly from the numerator and denominator, but belief about the meta-stable patterns *is the same*: an extrinsic 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)§. 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.

By relegating even the simple reference classes to the status of semantic conveniences (slobjects), aleatory and epistemic uncertainty are united in

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

§ The resulting philosophy of consciousness is left as an exercise to the reader.

miserably fumbling in the dark for answers. Sounds about right to a practising statistician.

> *Humanity knows nothing at all. There is no intrinsic value in anything, and every action is a futile, meaningless effort.* — Masanobu Fukuoka, quoted in Odell[26]

This is not cause for despair, but cause for humility and a rededication of effort and energy to goals of maintenance and caregiving. More on this in Section 5.1.

3.6 Aboutness, estimands and possible-worlds

I also find Yablo's work on "aboutness" useful.[29] 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.

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-sample MLE, but it will not hold for subjective, informative priors. So, I conclude that the frequentist and Bayesian (and other[30]) 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 for aleatory uncertainty.*

Aboutness matters for the fundamental and semi-mythical difference between Bayes and frequentism. $P(\ell \leq \theta \leq u)$ 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".[29]

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.

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

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, traditional frequentist teaching then defines confidence intervals in terms of frequencies over repeated data collections, 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.

We can think of the possible-worlds as having a common true θ and the nidcops as producing different $\hat{\theta}$s each time (Figures 1 and 2). But there might in a different conception of sampling space be multiple θs too (Figures 3 and 4).

3.7 Justified eclecticism

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 undermin-

ing the mathematical validity of what follows. My view, in agreement with Spiegelhalter, is that there is "no such thing as probability".[28]

That is, it is only a mathematical tool to optimise predictions and inferences, 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 performativity in science than it does about mathematics.

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, because they use a model where a reference class can be defined with sharp borders, though this is not really the case. That is not to say that Bayes is the more complete way to infer. We can choose different meanings for different purposes.

I set out my approach to using different statistical methods for different tasks in the table below. First, I reject the division of tasks into aleatory or aleatory+epistemic, for reasons of aboutness in Section 3.6. 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 with not a little discomfort, I will just call these *uninformative*) or beliefs about a data-generating process (which involves more informative or subjective priors).

I don't like this "informative" term, but I can't think of a better one without getting obscure.

Belief about the DGP does not 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 analyses in these two categories 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. 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-Wald tests, associated confidence intervals, Bayes factors with uninformative priors	Bayes factors with informative priors
Evaluate then identify (*a posteriori*)	MLE, associated confidence intervals; MCMC, ABC, *vel sim.*, with uninformative priors	MCMC, ABC, *vel sim.*, with informative priors

These are just examples of methods, and there are some situations that might place them differently; for example, the *a priori* testing methods of Neyman-Pearson-Wald 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).

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.

It would be natural to object to my classification of diffuse and similar priors on the same side as MLE and tests. After, all it is unfamiliar and somewhat dangerous to one's reputation to take an odd view like this. Few statisticians — perhaps none — have delved this far into ontological foundations of their work, and even then, many would feel constrained by the need to fit in and keep paying the bills. I don't mind that, I quite understand, and after all, the statistical analyses we do are built on supposition and simplification, as suggested in Figure 5.

I believe that I've justified having various philosophical stances on the nature of probability, but is it possible to be statistically eclectic while holding a *consistent* view? Having a clear frequentist stance, appealing to possible worlds and accepting that the reference class is not defined without some ambiguity, does not eliminate subjectivity. What about the opposite: does holding a de Finetti style opinion of personal belief as probability preclude long run methods and even hypothetico-deductive procedures? The long run

can be seen as a manifestation of the sampling process for the aleatory unknowns and even as the joint variation of θ and $\hat{\theta}|\theta$ for the epistemic ones, provided you again work sceptically with a possible-worlds ontology. Can you use the *a priori* methods in the table? I think so, provided that you do not muddle up the interpretations of your outputs by describing them as p-values or the like. On this basis, you do not see me presenting Bayesian p-values or Bayes factors, because the scope for misunderstanding is too great, I think.

4 Models and reality; aboutness revisited

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 otherwise unexplained influence (statistical noise) 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.

Figure 6: Henri Poincaré

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, unlike Boyle's law. Related to this, if we fit multiple competing models to the data and find that the results do not "converge" (in some casual sense of the word) to a shared interpretation, then we ought to back off and admit that the subject is inherently a tricky one to reduce as we would like.

We need to be sceptical about models, going even beyond Box's famous aphorism. I would rather build several complicated models with sensitivity analysis here and there, rather than do one thing and shrug off the fact that it is not useful in a Limitations paragraph. Henri Poincaré was an

*Also, many settings involve looking for tiny signals. The astronomer or particle physicist will push their instruments to the very limit of tolerable signal:noise ratio, and the commercial recommender system engineer will want to recommend a product even for customers who fit almost no previous pattern.

early influence on my thinking; he held that the choice of model (he was most famously focussed on geometry, but it is all the same with statistical models) was often driven by social convention and convenience. You will not see me writing up my one model that I used in some very specialised setting as a methodological publication, and when I see that (it is ubiquitous in statistical peer-reviewed journals), I wonder what the point of that is.

From Poincaré, you can trace steps easily enough to Manski, Gigerenzer, Platt, Lipton, Leamer and so on. Apply it everywhere and you might end up a resemblance nominalist.

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.[29] 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. (The analyst should have sorted out agreement on this before designing their data collection and analysis, and again before communicating results.)

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"[17] *. 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

* "Ranges are for cattle! Give me a number," said Lyndon Johnson, to a room full of economic advisors, perhaps apocryphally.

realism of the data-generating process, which are likely to be "marginal" and hence overlooked.

In the paper with Rick Hood,[3] 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,[31] 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, and some of them accommodate a sugar tax-mediated reduction in sugar consumption *and* an overall increase in sugar consumption via other sources.

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.[13] Poincaré also argued that the choice is often based on convenience and convention, and ought to prefer those explanations that explain the most phenomena (which Lipton called lovely explanations). From this, I draw more support both for flexible models and for communication that puts the audience first.

4.1 Latent variables and structural equation models

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 (one value for each group of observations), and so are imputed missing data (one value for each observation, but only some of them). But some of the most valuable applications come when we infer latent variables that take different values for every observation. To make these identifiable, we must 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[14]) 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.[33]

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.

5 Concluding preferences and the work environment

At this stage in the document, I need hardly say that I think the professional statistician must work against the prevailing hype in data analysis. We may have passed peak hype for big data analytics and machine learning, or it may be that a period of economic uncertainty has trimmed budgets for more speculative work (software development budgets are still growing, in 2023[34]), but the danger remains that unrealistic expectations, magical thinking, and a high-churn workplace will continue to undermine professionalism.

I wrap up in this section with the bits of the standpoint that are not justified by some grand philosophy but just by my own preferences. I don't attempt to throw every whim in here, like that I have a split keyboard and a trackpad in the middle, just those that link to what was previously covered and extend it.

5.1 Humans-in-the-loop; community; doing nothing

AI seems likely to automate the parts of statistical work that do not interest me so much: the feature engineering, the data cleaning, the running of many competing models and evaluation of sensitivity analyses. Yet the parts that are most fascinating — the definition of the question, understanding of the audience and effective communication of the results, not to mention

all that curation and compromise from Section 2.3 — remain the preserve of careful human work, which succeeds to the extent that there are effective interpersonal skills. Strange as it may sound:

Statistics is an interpersonal skill.

In some (many) unfortunate organisations, contemporary commercial data science valorises fast answers. The archetype is that the boss, asked a tricky question in the board meeting, can message the team downstairs and tell them to prepare a new model and report back with stats on its improvements, so that the boss can brag about it *before the meeting ends*. This is not a way to do anything useful. Of course, if the boss and the team all intend to churn, the board may not find out that it is nonsense until it's too late (at which point the new boss will be doing the same thing all over again). It's a depressing picture. I would prefer the conversation, the building of a complete picture of a system and what is really needed for business decisions, the step-by-step refinement.

I summarise this all by saying that I like to work on interesting problems with sensible people. Of course, I say this to people who like hearing it; I don't bother talking to the others. Sensible people are those who do not chase fashions, and who are willing to open up the problem and explore it with a professional. Otherwise, count me out.

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 why. 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 in the era of the "attention economy".[26] 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.

An important component is that my work ought to be physically located in and around a place. That's if you don't want to be worked into the ground travelling and hustling and trying to be all things to all people. It connects to the idea of being a personal brand. I tried that but now I describe myself as a statistician from Winchester, settling into a professional identity and a place, to do fewer things but do them well. I don't have time to be a well-oiled PR machine as well.

Remember Marc Augé? His definition of a place is the culture, including terms of interaction, established by those people in that physical location.[16] Here, I can make some contribution to building a *place* for statistical inquiry. It is about people, not media, not performance. This kind of detailed, discursive, non-performative work is what I mean by caregiving and maintenance work.

6 Historical opinions

The rewriting of this document from v1.4.3 into v2.0.1 involved decluttering, removing repetition and cutting out most of the jokey bits that had come to emulate that certain kind of philosophy writing where there are silly neologisms and wacky exemplaria. The novel terms that remain are essential central ideas — justified eclecticism (JE) and sampling space (rather sober terms, I hope) or a few shorthand terms that are deliberately not used consistently, like FSCs and nidcops. The section on aboutness was reduced to the essentials and philosophy of consciousness became a mere footnote.

After I learnt about Bayesian methods, I rapidly adopted them across most of my work, and became dismissive of frequentism. That was the time of growing anti-p-value sentiment and much lingering animosity could still be found in universities between the two "sides". In practice, this was tempered by the desire to serve the audience, and so rather than fit everything into a Bayesian mould, I always considered testing to be the right thing to do under a very limited set of circumstances (such as a randomised experiment where non-equivalence really was the question). Later, as I learnt more about philosophy of science, and as Deborah Mayo's critique of (insevere) Bayes took shape, I started building the careful basis for my views which you see here, and that made my stance more nuanced but also more secure.

As detailed in Section 3.7, I no longer choose to divide statistical inferential methods into Bayesian and non-Bayesian. But, as that is the term widely used and understood, I would still describe myself as a statistician specialising in Bayesian models, etc. I run "Bayesian" courses, and so forth.

Throughout, I have been critical of the marginalising language games such as credible, not confidence interval, or prediction, not estimation of group-level effects. There are times when I have used the term "confidence interval" for a Bayesian 95% percentile central interval from the marginal posterior sample.[32] This particular paper also features a (successful, I think) attempt to define a Bayesian significance threshold — though we did not use the term significance:

> *"As the Bayesian model does not provide P-values in the traditional sense, a meaningful difference was set at* $> 1\%$ *attainment score and when the confidence interval did not cross 0."*

7 To-do list

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

- is there always a contentious zone in sampling space?
- what minimal version of nominalism is required to comfortably imply justified eclecticism?
- Carnap's logical probability and induction
- persistence over time
- Keynes' weight of evidence — most recently revived by Margherita Harris — and links (maybe) to Manski's uncertainty
- am I saying that simple things exist and accumulate into FSCs (page 23)? or that hard-to-predict patterns accumulate into easy-to-predict patterns, at least sometimes (page 33)? or both, or neither?

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 1.4.3 to 2.0.1). Substantive additions are recorded with a minor version number (e.g. from 1.3.2 to 1.4.1). 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).

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] 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.

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

[5] Bååth R. *Introduction to Bayesian data analysis — Part 2: Why use Bayes?* https://www.youtube.com/watch?v=mAUwjSo5TJE&list=PLvrRa_pKW100YldDeD660XakB7lm8Rwqn&index=2 Accessed 15 February 2023.

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

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

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

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

[10] Dodzo W, Grant RL, Forsyth L, Ramdharry GM. A randomised controlled feasibility trial of the Graded Repetitive Arm Strengthening Programme (GRASP) delivered to stroke survivors at home. *International Journal of Therapy and Rehabilitation* (2020); 27(8). doi: 10.12968/ijtr.2017.0081

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

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

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

[14] 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/

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

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

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

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

[19] Hájek A. Fifteen Arguments Against Hypothetical Frequentism. *Erkenntnis* (2009); 70: 211–35. doi: 10.1007/s10670-009-9154-1.

[20] 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.

[21] The Royal Burgh of Pittenweem and District Community Council. *Pittenweem Community Survey 2012*, p.3. https://www.aboutpittenweem.org.uk/uploads/1/8/2/9/18297767/pittenweem_community_survey_2012.pdf Accessed 8 February 2023.

[22] 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/ Accessed 8 February 2023.

[23] 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/ Accessed 8 February 2023.

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

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

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

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

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

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

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

[31] 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.

[32] Norris M, Hammond JA, Williams A, Grant R, Naylor S, Rozario C. Individual student characteristics and attainment in pre-registration physiotherapy: a retrospective multi site cohort study. *Physiotherapy* (2018); 104: 446–52.

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

[34] Lohr S. Beyond Silicon Valley, spending on technology is resilient. *New York Times*, 13 February 2023. https://www.nytimes.com/2023/02/13/technology/ technology-spending-resilient.html