Frequentism as a positivism: a three-tiered interpretation of probability

Shivaram Lingamneni

July 29, 2013

Abstract

I explore an alternate clarification of the idea of frequency probability, called *frequency judgment*. I then distinguish three distinct senses of probability — physical chance, frequency judgment, and subjective credence — and propose that they have a hierarchical relationship. Finally, I claim that this three-tiered view can dissolve various paradoxes associated with the interpretation of probability.

1 Introduction

1.1 Frequentism and its challenges

Frequentism means, more or less, that probabilities are ratios of successes to trials. It originates with John Venn and is arguably the first philosophically rigorous account of probability — that is to say, it is the first account of probability to appear as an attempt to correct a philosophically inadequate pre-theoretic view. As Alan Hájek has observed, however, it has fallen on hard times. In part, this is because it competes with the Bayesian interpretation of probability, in which probabilities are subjective degrees of belief. Bayesianism offers a seductive unifying picture, in which epistemology and decision theory can both be grounded in a quantitatively precise account of an agent's attitudes and propensities. But frequentism's philosophical difficulties are not simply due to its being outshone by a competing view. As Hájek has shown, frequentism itself faces a variety of vexing challenges.

Hájek reconstructs frequentism as containing two distinct conceptions of probability — *finite frequentism*, in which probabilities are actual real-world ratios of successes to trials, and *hypothetical frequentism*, in which they are limiting relative frequencies over an idealized hypothetical infinite sequence of trials. In a series of two papers [1996, 2009], he shows that each conception

is affected by numerous difficulties: in fact, each paper gives 15 distinct objections to one of the conceptions!

In order to motivate what follows, I'll briefly summarize what I consider the most pressing of Hájek's objections against each characterization. Finite frequentism is intuitively appealing because of its metaphysical parsimony; probabilities can be "read off" from the actual history of real-world events, without the need to posit any unobservable entities. But taken literally, it clashes with many of our important intuitions about probability. In particular, it is a kind of *operationalism* about probability, and hence suffers from similar problems to other operationalisms. If we consider probability to be defined by real-world frequency, then we have seemingly have no way to express the idea that an observed frequency might be aberrant, just as defining temperature to be thermometer readings leaves us with no way to express the idea that our thermometers may be inaccurate. This problem becomes especially serious when we consider cases where the number of real-world trials is very small — in particular, if there is only 1 trial, then the finite frequency probability must be either 0 or 1, and if there have been no trials yet, then it is undefined. Finite frequentism is in conflict with our intuitions that actual trials constitute *evidence* about probability rather than its actual substance.

Hypothetical frequentism answers this concern perfectly, but at far too high a metaphysical cost. In particular, asserting the existence of an infinite sequence of trials seems to involve an "abandonment of empiricism." In the real world, we cannot perform an infinite sequence of trials, so the meaning ascribed to probabilities is evidently counterfactual. Even after granting this, what kind of counterfactual are we dealing with? If we analyze it using a possible-world semantics, in the style of Stalnaker or Lewis, we seemingly require a possible world that (at the very least) violates the conservation of mass-energy. Why should we believe that probabilities in this world have anything to do with ours?

Finally, the following objection is commonly advanced against both conceptions of frequentism: frequentism entangles the probability of any individual event E with the question of what will happen to other, similar events. We cannot make frequentist sense of the probability of E without assigning it to some broader *reference class* of events, over which we will be able to define a ratio of successes to trials. But at this point, $P(E)$ will be a property of the reference class, not of E itself. This objection is already troubling, but it has even more teeth in cases when there are multiple possible reference classes, each yielding a distinct value of $P(E)$, or perhaps no reference class at all. This is the so-called "reference class problem", and it is another, crucial sense in which frequency notions of probability diverge from our ordinary understanding of the word.

1.2 Where to?

I am a frequentist. What sort of frequentist am I? Of the two varieties distinguished above, I am much more sympathetic to finite frequentism; the metaphysical costs of infinite hypothetical sequences are too much for me to bear. In fact, I think that finite frequentism, properly expounded, can actually escape many of the criticisms Hájek levels at it — perhaps eight out of fifteen. But I cannot deny the force of Hájek's overall arguments, and I think it inevitable that I must give some ground. Specifically, I think an adequate analysis of probability must both seek a third way of defining frequency probability and also acknowledge that not all probabilities are frequency probabilities. Here are some of my desiderata for such an expanded conception:

1. It should preserve core frequentist intuitions that relative frequency is an essential component of probability. In particular, it should not conflate probabilities that have an intuitively acceptable frequency interpretation (e.g., the probability that a U.S. quarter, when flipped, will land heads) with those that do not (e.g., the probability referenced in Pascal's wager that God exists).

   Indeed, the primary goal of this paper is to propose and defend a definition of frequency probability that is both reasonably rigorous and free from paradox, in hopes that it will enable epistemological views in which frequency probability has a privileged status.

2. It should not take a stance on the existence of physical chance (something which poses problems for both frequentist and Bayesian accounts of probability). I think that a proper resolution of this question rests on questions external to the philosophy of probability, in particular on the philosophy of physics, and that consequently it is an advantage for an account of probability to remain agnostic on the question.

3. It should not deny the validity of the Bayesian interpretation of probability outright. As Jaynes [1985] remarked, arguing in the reverse direction:

   > I do not "disallow the possibility" of the frequency interpretation. Indeed, since that interpretation exists, it would be rather hard for anyone to deny the possibility of it. I do, however, deny the necessity of it.

   Indeed, while I consider myself a frequentist, I affirm the value of Bayesian probability, both its technical validity as a consistent interpretation of the laws of probability and as the correct solution to certain

epistemological problems such as the preface paradox. My skepticism is confined to claims such as the following: all probabilities are Bayesian probabilities, all knowledge is Bayesian credence, and all learning is Bayesian conditionalization. I will say more about this later.

4. At the level of statistical practice, it should support a methodological reconciliation between frequentist and Bayesian techniques. That is to say, it should acknowledge that in practice both methods are effective on different problems, independently of the philosophical debate. Kass [2011] calls this viewpoint "statistical eclecticism" and Senn [2011] calls it "statistical pragmatism".

5. Thus, it is necessary for it to preserve the distinction between frequentist and Bayesian methods, that is to say, between methods that make use only of probabilities that have a natural frequency interpretation and those which make use of prior probabilities that do not. Otherwise, frequentist and Bayesian methods are collapsed into a single group, in which frequentist methods appear merely as oddly restricted Bayesian methods.

Without further ado, I will introduce an account of probability that I believe will fulfill all these criteria. The argument will necessarily detour through many philosophical considerations related to probability. The reader who is pressed for time should look at sections 2, 4, and 5.

1.3 Precedents for the view

The closest historical precedent I am aware of for my view is Carnap's distinction [1945] between two senses of probability: Probability$_1$, which describes credence or degree of confirmation, and Probability$_2$, which describes long-run relative frequency over a sequence of trials. In particular, he makes the following parenthetical remark about Probability$_2$:

> I think that, in a sense, the statement '$c(h, e) = \frac{2}{3}$' itself may be interpreted as stating such an estimate; it says the same as: "The best estimate on the evidence e of the probability$_2$ of M_2 with respect to M_1 is 2/3." If somebody should like to call this a frequency interpretation of probability, I should have no objection.

My view differs substantially from Carnap's in almost all respects — in particular, I will not make use of the notion of *logical probability* that he advocated. Nevertheless, I will interpret this remark as Carnap's blessing.

2 The theory

Three conceptually distinct interpretations of probability suffice to describe all uses of probability. They are arranged in a tiered hierarchy as follows:

1. Physical chance, if it exists. This is the only objective and metaphysically real kind of probability.

2. Frequency judgments. Pending a more precise motivation and definition, the core idea is this: given an event E, a frequency judgment for E is a subjective estimate of the proportion of times E will occur over an arbitrarily large (but finite) sequence of repeated trials. This is intended as a frequency interpretation of probability, i.e., one that can replace finite and hypothetical frequentism.

3. Bayesian subjective probability in the sense of Ramsey and de Finetti.

Probabilities pass "downwards" along this hierarchy in the following sense:

1. If an agent knows a physical chance (and no other relevant information), that agent is obliged to have a frequency judgment coinciding with the physical chance.

2. If an agent has a frequency judgment (and no other relevant information), that agent is obliged to have a Bayesian subjective probability coinciding with the frequency judgment.

Thus, as we pass down the hierarchy, the domain of applicability of the interpretations strictly increases. In particular, the conjunction of the two relations yields a large fragment of (possibly all of) Lewis's Principal Principle.

3 The first tier: physical chance

Lewis [1994] defines chance as "objective single-case probability", which does an excellent job of explaining why chance is so vexing for both frequentists and Bayesians. For one, a chance is a probability that we intuit as being objectively real, which is at odds with radical Bayesian subjectivist accounts in which all probabilities are agent-relative and have to do with dispositions to act. Thus, it is typical for Bayesians to accept chances, when they exist, as an additional constraint on belief beyond that of simple consistency, in the form of Lewis's Principal Principle. This principle has varying formulations, but the rough idea is that if an agent knows the chance of an event E, and

they have no other relevant information, they should set their credence in E to be the same as the chance.

But chance is also problematic for frequentists because of the intuition that they exist in the *single case* — a chance seems no less real despite only being instantiated once, or perhaps not at all. Lewis gives the memorable example of unobtainium, a radioactive heavy element that does not occur in nature, but can only be produced in a laboratory. One of the isotopes, Unobtainium-366, will only be instantiated twice as atoms. The other, Unobtainium-369, will never be instantiated at all (perhaps due to budget cuts). In the case of Unobtainium-366, we intuit that the true half-life of the isotope (phrased equivalently in terms of probabilities, the objective chance of decay within a particular fixed time period) may be something quite different from anything we might generalize from our two observed data points. In the case of the heavier isotope, we have no data points at all to go on. So there is a conflict with any frequentism that insists that probabilities are always synonymous with actual frequencies, or can always be straightforwardly extrapolated from them.

But this is not yet the whole story about why chance is problematic. There are two rather different senses in which physical chance appears in accounts of probability. One is the existence of physical theories, for example the Copenhagen and objective collapse interpretations of quantum mechanics, in which reality itself is nondeterministic and thus the existence of chances is a physical and metaphysical fact about the universe. But the other is when a physical phenomenon appears, on empirical grounds, to have irreducibly probabilistic behavior. Radioactive decay is one example, but another particularly intriguing case, appearing in Hoefer [2007] and Glynn [2010], is Mendelian genetics, e.g., the probability that two carriers of a recessive gene will have a child in whom the gene is expressed.

Thus we encounter a dispute in the literature: is the existence of physical chance compatible with a deterministic universe? One intuitive answer is no: if the course of events is determined, then chance is annihilated and the chance of any individual event E is 1 if it deterministically occurs and 0 if it does not. This was the view of Popper and Lewis and it has continuing defenders, in particular Schaffer [2007].

However, other authors defend the idea that a deterministic universe could exhibit chance. For example, Lewis wanted chance to supervene (in a Humean sense) on past, present, and future spatiotemporal events, rather than existing as a distinct metaphysical property. He accomplished this via the so-called "best-system analysis", on which considerations such as symmetry or extrapolations from related systems can be chancemakers beyond mere sequences of events. Although Lewis himself believed chance to be incompatible with determinism, nothing about such an analysis requires indeterminism

and it can support a compatibilist account of chance, as in Hoefer and Eagle [2011]. Glynn also defends deterministic chance, but he is motivated instead by the existence of probabilistic scientific laws, such as Mendelian genetics or statistical mechanics, that would hold even in a deterministic universe. Thus, he is essentially making an indispensability argument; if chance is essential to our understanding of the laws of Nature, then we are not justified in denying its existence due to metaphysical qualms.

It follows that the question of whether chance exists is undecided. If you believe the Copenhagen interpretation of quantum mechanics, then measuring a quantum superposition such as $\frac{\sqrt{2}}{2}(|0\rangle + |1\rangle)$ yields either 0 or 1, each with probability $\frac{1}{2}$, and the outcome is not determined in any sense before the measurement. This is then a source of objective randomness and fulfills the criteria for physical chance. If you are undecided about quantum mechanics, but believe Glynn's arguments about chances from laws, then there is still an objective chance of whether two heterozygous parents will have a homozygous child. But if you believe the de Broglie-Bohm interpretation of quantum mechanics, in which reality is deterministic, and you also endorse Schaffer's denial of deterministic chance, then there are no nontrivial physical chances.

My purpose in proposing physical chance as the "highest" interpretation of probability is not to adjudicate the question of whether chance exists, and if so, what exactly it is.[1] Rather, I am offering people with different views of chance a blank check which they can fill in with their preferred conception. The proper interpretation of quantum mechanics is a question for physicists and philosophers of physics; whether Glynn's argument is correct seems to hinge, like other indispensability arguments, on deep questions about whether scientific practice justifies scientific realism. Separating chance from other notions of probability lets us separate these questions from the debate about what probability itself means.

4 The second tier: frequency judgments

My characterization of frequency probabilities will rest on two primitive notions. One is that of a reference class: a reference class is simply a description that picks out a class of events. In the typical case, a reference class

[1] In passing, I do have some sympathy towards the idea of deterministic chance, in particular for microphysical events. For example, measuring $\frac{\sqrt{2}}{2}(|0\rangle + |1\rangle)$ produces an apparently random sequence of 0s and 1s, no matter what interpretation of quantum mechanics one favors, and there seems to be a fine case for such a phenomenon exhibiting chance. I become increasingly skeptical as this argument is extended upwards to macrophysical phenomena, such as genetics. I am also unimpressed with the best-system analysis as such, which strikes me as a confusion of metaphysics with epistemology. But this is a digression from my main argument.

will preferably satisfy some other criteria, for example Salmon's [1971] notion of homogeneity: that there is no additional criterion, or "place selection function", that picks out a subclass with substantially different properties. However, my discussion here will not impose any such additional requirements. One of the strengths of probabilistic analysis is that it can be applied to data that are not "genuinely random" in any meaningful sense — in an extreme but instructive case, the output of a deterministic pseudorandom number generator. If the analyst considers the data to defy a deterministic analysis, or just that they can benefit from a probabilistic one, that is sufficient.

The second primitive notion is that of *epistemically* independent events; this is a kind of pre-theoretic counterpart to the idea of mutual independence. Events are epistemically independent when knowing the outcome of some does not does not tell us anything useful about the outcome of any other. This is a subjective notion relative to the agent's knowledge and needs; in particular it is not necessary that the events, should they have objective chances, have probabilistically mutually independent chances, or that the agent take into account all available evidence about how the events might be related.

Definition 1. *Given an event E and a reference class R for it, an agent A's frequency judgment for E is a real number $p \in [0, 1]$, representing a subjective estimate of the proportion of times E will occur over an arbitrarily large (but finite) sequence of epistemically independent trials in the chosen reference class R.*

Having a frequency judgment of p for E is a sufficient condition to model E as being drawn I.I.D. (independently and identically distributed) from the Bernoulli distribution with parameter p. That is to say, in intuitive terms, we can model E in the same way as we would model flips of a coin with bias p. This is not to say that we model E *as* such a coin — this would be a circularity, since we need the definition of frequency judgment to clarify what it means for the coin to have long-run bias! Rather, each situation has a natural representation as a Kolmogorov-consistent probabilistic model, and the resulting models are in fact the same.

In order for estimates of this kind to make sense, we require a clear conception of the reference class R supporting an arbitrarily large number of trials. The motivation for this is clear: we can toss a coin an arbitrary number of times to clarify the relative frequency of heads, but we cannot repeat a one-off event such as the 2000 U.S. presidential election to examine any probabilistic variability in its results. Looking back to our discussion of chance, all the chance-like physical phenomena we discussed (quantum measurements, radioactive decay, and Mendelian genetics) admit frequency judgments, even if they are excluded by a specific account of chance. Even

the decay of Unobtainium-369, the element that will never be instantiated, admits one because we have a clear and unambiguous conception of what it would mean to synthesize its atoms and measure the incidence of decay. Thus, the existence of this intermediate interpretation of probability — less objective than physical chance, but more so than Bayesian credence — should soften the blow of deciding that some chance-like phenomena do not genuinely exhibit chance.

4.1 Invariance under averaging

There are some formal difficulties with the definition of frequency judgment. What does it mean to have a non-integer estimate of the number of times E will occur over a integer-long sequence of trials? And why, if frequency judgments are estimates of proportions over finite sequences, is it possible for them to take on irrational values?[2] I think the natural resolutions of these problems succeed, but it is not entirely obvious that they succeed honestly; one might suspect that they are parasitic on a prior, unexplained concept of probability or expected value. So I will give a brief argument to justify that real-valued proportions are sensible as frequency judgments.

The intuition is this. Consider someone who can give integer-valued estimates of the number of successes over n trials, for arbitrary n. We ask him for his estimate of the number of successes over a single trial, and he tells us either 0 or 1. Now we ask him, "if you repeated that single trial 10 times, then averaged the number of successes over the 10 repetitions, what would you estimate the average to be?" Because epistemic independence implies that there is no difference between a 10-trial block and 10 1-trial blocks, he should give us his estimate of the number of successes over 10 trials, divided by 10: this will be the first decimal digit of his real-valued frequency judgment. We can continue this process to elicit more digits, or we can simply ask him to "tell us the averages first," rather than bothering with the integer estimates. Formally:

Definition 2. *Given an event E and a reference class R for it, an agent A's frequency judgment scheme for E is a map $f : \mathbb{N} \to \mathbb{R}$, such that $f(n)$ is a subjective estimate of the number of times E will occur over n epistemically independent trials of R. Evidently, $f(n) \in [0, n]$ for every n.*

So at this point, we are considering both frequency judgments in the original sense, but also schemes that make integer predictions for every n.

[2]This is Hájek's 14th criticism of finite frequentism. There I think it succeeds to some extent — unlike frequency judgments, there is an essential sense in which actual frequencies are rational numbers. Of course, one could argue for the use of real numbers there too, as an idealizing assumption that enables the use of continuous mathematics.

But now we impose another criterion: f should be *invariant under averaging*. In other words, let us say that f estimates that if we do n trials, we will have s successes. We should also estimate that if we do $2n$ trials and then divide the number of successes by 2, we should get s. In other words, we should have $\frac{f(2n)}{2} = f(n)$.

In general, for any $a \in \mathbb{N}$, our estimate should be invariant under averaging over a repetitions of the trial, i.e., $\frac{f(an)}{a} = f(n)$. But this implies that f should satisfy $f(an) = af(n)$ for any $a \in \mathbb{N}$. Now, fix some n and let $p = \frac{f(n)}{n}$; clearly p is a real number in $[0, 1]$. For any $m \in \mathbb{N}$, $nf(m) = f(mn) = mf(n) = mpn$. Dividing by n, we get that $f(m) = pm$ for all m. We have shown that frequency judgment schemes that are invariant under averaging are necessarily frequency judgments, i.e., real-valued proportions.

Mathematically speaking, this argument is trivial; its significance is that we appealed only to a notion of averaging over arbitrary repetitions, without any circular appeal to probability or expected value. Furthermore, I think this argument yields two important clarifications of the idea of frequency judgment:

1. The concept of invariance under averaging gives rise to a simple notion of "long-run relative frequency" without appealing to an infinite sequence of trials. Thus the frequency judgments interpretation appropriates some of the benefits of hypothetical frequentism as analyzed by Hájek, without having to carry any of its metaphysical baggage.

2. If f is invariant under averaging, then $f(n) = nf(1)$. Thus, in some sense f "views" every individual trial as contributing a fractional success $f(1) \in [0, 1]$ to the total estimate of successes. This is what justifies modeling events that admit a frequency judgment as I.I.D. Bernoulli trials.

A concern remains: why is it sensible for p to take on irrational values? The key is that the reals are Archimedean, i.e., for any two reals r_1, r_2, we have $|r_1 - r_2| > q$ for some rational q. It follows that over a sufficiently large integer number of trials, any two distinct reals constitute distinguishable frequency judgments; their estimates of the number of successes will vary by at least one whole trial. For example, consider the irrational-valued frequency judgment $\frac{\pi}{4} \approx .785398$. Is this judgment identifiable with any rational-valued approximation of it, e.g., .785? It is not, because over 100000 trials, they predict quite different things.

At this point, one might take issue with the idea that arbitrary-precision real numbers are distinguishable in this way. Surely, at some point, the number of trials required to make the distinction is so large that the heat

death of the universe will come first? I appreciate this concern, but I don't think it's specific to probability — it seems akin to the idea that instead of modeling time as real-valued quantities of seconds, we should model it as integer multiples of the Planck time. There may be a bound on the resolution of reality, but it is methodologically convenient to represent it as unbounded.

5 Characteristics of frequency judgments

5.1 Caveats

It is problematic to claim that frequency judgments are in fact a frequency interpretation of probability, and I do not wish to paper over the difficulties. This conception is a substantial retreat from the classical frequentism of Reichenbach and von Mises. In particular:

1. A frequency judgement is not "made by the world"; it is not directly derivable from any actual past history of trials (as in the case of finite frequentism), the past and future history of the world (as in some cases of Lewis's supervenience account), or any objective or universal conception of an idealized hypothetical sequence of trials (as in the analogous case of hypothetical frequentism).

2. A frequency judgment is explicitly relative to both an agent, because it is a subjective estimate, and to a reference class. These relativizations may look like reluctant concessions to realism, but in my opinion they are features, not bugs — they capture essential indeterminacies that must be part of any positivist account of probability. I will say more about both relativizations below.

3. A frequency judgment need not pertain to events that are truly "random" in any sense. Deterministic phenomena that are too difficult to analyze with deterministic methods (such as the operation of a pseudorandom number generator), when analyzed probabilistically, can be classed at this level of the hierarchy. Thus, von Mises's analysis of randomness by means of the notion of Kollectiv (an idealized infinite random sequence with certain desirable mathematical properties) is not relevant.

4. The notion of frequency judgment is intended as a conceptual analysis of probability — it is an attempted elucidation of what is meant by statements such as "the probability of flipping a U.S. quarter and getting heads is $\frac{1}{2}$," or "the probability of a Carbon-14 atom decaying in 5715 years is $\frac{1}{2}$." It does not follow from this that an agent's frequency

judgments are necessarily a completed totality and form a σ-algebra obeying the Kolmogorov axioms.

A frequency judgment is not necessarily part of any global probability distribution, even one relative to a particular agent; it is created by an act of model-building and can be revised arbitrarily in ways that do not correspond to conditional update.

5.2 Relativization to reference classes

Frequency judgments are explicitly relativized to reference classes. Does this mean that they cannot be an analysis of probability simpliciter? Concerning this question, I endorse the argument by Hájek [2007] that in fact, every interpretation of probability is affected by a reference class problem, and thus explicit relativization to reference classes is needed to dissolve an intrinsic ambiguity.

I will briefly sketch Hájek's argument as it applies to Bayesian subjective probability. According to the most radical accounts of subjective credence, there are no constraints on credence besides mere consistency. But intuitively, such a view is unsatisfying because it does not enforce any kind of relationship between one's beliefs and reality. Hájek gives the following memorable example:

> The epistemology is so spectacularly permissive that it sanctions opinions that we would normally call ridiculous. For example, you may assign probability 0.999 to George Bush turning into a prairie dog, provided that you assign 0.001 to this not being the case.

Thus it seems necessary to admit additional constraints on belief — for example, Lewis's Principal Principle, in which beliefs must coincide with known chances, or Hacking's Principle of Direct Probability, in which they must coincide with observed relative frequencies. But external "testimony" of this kind is, by its nature, subject to a reference class problem. Consider the following case: John is 60 years old, a nonsmoker, and previously worked with asbestos. We have statistics for the incidence of lung cancer in 60-year-old nonsmokers and 60-year-olds with asbestos exposure, but we have no statistically significant data concerning the intersection of those groups. What should our credence be that John will develop lung cancer? We might pick the first rate, or the second, or try to interpolate between them, but implicit in any of these decisions is a statement about what reference class is to be preferred.

Hájek's conclusion from this analysis is that we need new foundations for probability; he considers the true primitive notion of probability to be conditional probability, where the assignment of the event to a reference class

is part of the proposition being conditioned on. That is to say, instead of considering $P(A)$, Hájek thinks we should be looking at $P(A \mid A \in R)$, where R is a reference class. I think that the frequency judgments interpretation, in which the reference class is part of the definition of (unconditional) probability, is a more natural way of addressing this issue, and one that allows us to retain our existing foundations. I discuss this question further in section 10.3.

5.3 Relativization to agents

The frequency judgments interpretation makes no reference to infinite sequences or possible worlds; it relies only on the conceivability of performing additional representative trials. Thus, its closest relative in terms of metaphysical commitments is finite frequentism. But frequency judgments are quite unlike finite frequencies in that they are agent-relative; two different agents can have two different frequency judgments, even after they come to agreement about a reference class. I will try to motivate this with a simple case study. Consider the case of a coin that has been flipped 20 times and come up heads 13 times. Is an agent constrained, on the basis of this data, to have any *particular* estimate of the proportion of heads over a long sequence of trials? Intuitively, the answer is no; a variety of beliefs about the coin's long-run behavior seem perfectly well justified on the basis of the data.

The ambiguities in estimation begin with the reference class problem. One reading of finite frequentism is we must assign $P(H)$ to be $\frac{13}{20}$, the ratio of actual successes to actual trials. This could be quite reasonable in some circumstances, e.g., if the coin seems notably atypical in some way; however, to say that finite frequentism *requires* this value is to do it an injustice. A finite frequentist might also say that the reference class provided by the sample is deficient because of its small size, and choose instead the reference class of *all* coinflips, yielding a $P(H)$ of $\frac{1}{2}$, or rather, negligibly distant from $\frac{1}{2}$. But the spectrum of choices does not end there.

The maximum likelihood estimate of the probability of an event E is $\frac{s}{n}$, the ratio of successes to trials; this is a frequentist estimator in the sense that it does not involve the use of prior probabilities. As such, it coincides with the first reading of finite frequentism and estimates $P(H)$ to be $\frac{13}{20}$, but it would be a mistake to identify the two perspectives. Rather, the maximum likelihood estimate is the value of $P(H)$ under which the observed data are most probable; this is not an ontological attribution of probability but explicitly an estimate. As such, it competes with Bayesian estimators such as the Laplace rule of succession, which begins with a uniform prior distribution over the coin's biases and conditions repeatedly on each observed flip of the coin. The resulting posterior distribution is the beta distribution $\beta(s+1, n-s+1)$; to get the estimate of the posterior probability, we take

its expected value, which is $\frac{s+1}{n+2} = \frac{14}{22}$.

Since the rule of succession is derived from a uniform prior over the coin's biases, a different Bayesian might use a different prior. For example, using a prior that clusters most of the probability mass around $\frac{1}{2}$, such as $\beta(n, n)$ for large n, will produce an estimate arbitrarily close to $\frac{1}{2}$. But on a different note entirely, another frequentist might start with a null hypothesis that the coin is fair, i.e., $P(H) = \frac{1}{2}$, then compute the p-value of the observed data to be 0.26 and accept the null hypothesis, retaining the estimate of $\frac{1}{2}$.

None of these answers is prima facie unreasonable — even though they differ considerably in methods and assumptions, they are all legitimate attempts to answer the question, "if this coin is flipped a large number of times, what proportion of the flips will be heads?" I am therefore rejecting Carnap's suggestion that there should in general be a *best* estimate of long-run frequency from the data. We will have to live with a multiplicity of frequency judgments, because room must be left for legitimate differences of opinion on the basis of data.

5.4 Frequentism as a positivism

Given all this, why are frequency judgments still a frequency interpretation of probability? I think they preserve the content of frequentism in two important senses. First, their definition depends essentially on the notion of repeated trial. If there is no conception of a reference class of trials, then there can be no frequency judgment. Thus, frequency judgments reflect the intuition that there is no way to make frequentist sense of probability claims about one-off events.

More crucially, even though frequency judgments are not objective, they are directly falsifiable from empirical data. Consider the example in the previous section: on the basis of observing 13 heads over 20 trials, we considered a range of different frequency judgments about $P(H)$ to be valid. But no matter what value we chose, we have a clear conception of how to further clarify the question: we need to flip the coin more times and apply some statistical test that can differentiate between the different judgments.

For example, consider the case of someone whose frequency judgment for $P(H)$ is $\frac{2}{5}$. If we go on to flip the coin 1000 times and get 484 heads, then (using the normal approximation to the binomial) our observed result is 5.42 standard deviations from the mean of 400 heads predicted by their hypothesis, which yields a p-value on the order of 10^{-8}. This is so highly improbable that we may consider the frequency judgment of $\frac{2}{5}$ to have been falsified. This is not to say that the much-maligned p-value test is the gold standard for the falsification of frequency judgments; likelihood ratio tests can be used to achieve the same results. If two agents can consense on a reference class for E, they can settle whose frequency judgment for $P(E)$ is

correct.

This explains the intuition that frequency probabilities are objective. If there is a large, robust body of trials for an event E (such as coin flipping), then any frequency judgment that is not extremely close to the observed finite frequency is already falsified. Thus, for events such as "a flipped U.S. quarter will land heads", our expected frequency judgment (in this case $\frac{1}{2}$) is very nearly objective.

How essential are repeated trials to this idea of probabilistic falsification? Indeed, it is possible for a Bayesian probability for a one-off event to be falsified, in the cases when that probability is very large or very small. For example, if an agent makes the subjective probability assignment $P(E) = .00001$, and then E in fact comes to pass, then the agent's assignment has been falsified in much the same sense as we discussed above. But if E is one-off, an credence like $P(E) = 0.5$ that is far away from any acceptable threshold of significance cannot be falsified. The event E will either occur or fail to occur, but neither of these will be statistically significant. Such a Bayesian credence lacks any empirical content.

In this sense, the definition of frequency judgment is an attempt to recover the purely *positivist* content of frequentism. The metaphysical aspect of frequentism, in which probabilities are inherently real and objective, has been deferred to the level of chance. Inasmuch as Bayesian credences are purely matters of personal opinion, without empirical content, they are also deferred to another level.

5.5 Calibration

To remedy this, the literature on Bayesianism proposes the notion of *calibration*: a Bayesian agent is calibrated if $\frac{1}{2}$ of the events he assigns credence $\frac{1}{2}$ to come to pass, and so on.[3] Calibration does seem to restore empirical content to single-case subjective probability assertions — intuitively, given a one-off event E, a subjective declaration that $P(E) = \frac{1}{2}$ is more empirically justified coming from an agent with a strong history of calibration than from one without one. The problem is that calibration, as a norm on subjective agents, represents a substantial compromise of the Bayesian view, so much so that it cannot be taken to save the original notion of subjective probability from these criticisms.

Firstly, as Seidenfeld [1985] observes, calibration is straightforwardly dependent on a notion of frequency probability, and what that notion is requires explication. In what sense are we to interpret the statement that $\frac{1}{2}$ of the events will come to pass? Seidenfeld considers finite-frequentist ("$\frac{1}{2}$ of these

[3]To solve the problem of sparseness, it is common to discretize or "bucket" the credences, e.g., by including also the events which were assigned credences in $[.45, .55]$.

events have historically come to pass") and hypothetical-frequentist ("the long-run relative frequency of these events coming to pass is $\frac{1}{2}$") readings of this claim and rejects them, for reasons akin to the difficulties Hájek sees with these interpretations in general.[4]

Can we make sense of calibration under the tiered interpretation? In fact, an assertion of calibration has a straightforward interpretation as a frequency judgment: the agent is taking the class of events she assigns subjective probability $\frac{1}{2}$ to be a reference class, and then making a frequency judgment of $\frac{1}{2}$ for that class. This is an empirical assertion, subject to confirmation or disconfirmation in the manner discussed in the previous section. However, this notion of confirmation is a property not of the single case, but of the class of predictions as a whole.

Secondly, just as calibration inherits the problems of definition that affect frequency probability, it also inherits a reference class problem. For example, van Fraassen [1983] gives the following surefire technique to achieve calibration: make 10 predictions, on any subject, with probability $\frac{1}{6}$. Then, roll a fair die 1000 times, predicting an outcome of 1 each time with probability $\frac{1}{6}$. At the end of this, you will (with high objective probability) be calibrated, in the sense that almost exactly $\frac{1}{6}$ of your predictions with probability $\frac{1}{6}$ will have come true. But clearly your ability to make calibrated predictions about the die says nothing about your predictive ability in general — it is unreasonable to place the original 10 predictions and the subsequent 1000 in the same reference class.

Both of these difficulties have a common theme: calibration, as a norm, entangles individual subjective probability assertions with the facts about a larger class of events. Thus it cannot be taken to provide empirical content for single-case probability assertions. And inasmuch as this empirical content does in fact exist, my claim is that it is captured exactly by the notion of frequency judgment: it is no more and no less than the ability to define an arbitrary reference class and make a relative frequency assertion about it.

6 The third tier: Bayesian probability

I will use the term "probabilism" to describe the following view:

[4]Seidenfeld also cites a theorem by Dawid [1982], which asserts that according to a Bayesian agent's own subjective probability distribution, she will necessarily achieve long-run calibration with probability 1. This is a consequence of the Law of Large Numbers — compare the observation that if a coin is in fact fair, even after an initial sequence of 999 heads and 1 tail, the relative frequency of heads will still converge in the limit to $\frac{1}{2}$ with probability 1. Dawid and Seidenfeld take this to mean that the idea of calibration is either trivialized or inexpressible under a strict Bayesian interpretation of probability. But see new work by Sherrilyn Roush for an account of Bayesianism in which calibration is a nontrivial norm on subjective probability.

1. Uncertain knowledge and belief can (at least some of the time) be modeled by probabilities ("credences", "Bayesian personal probabilities").

2. These credences can (at least some of the time) be measured by an agent's disposition to act or bet.

3. Credences ideally satisfy the Kolmogorov axioms of probabilistic consistency.

4. The desirability of this consistency is demonstrated by the Dutch Book argument.

According to this definition, I consider myself a probabilist. It seems perverse to me to try and dispense entirely with the idea of real-valued credence — at the very least, I really do have propensities to bet on a variety of uncertain events that have no frequency interpretation, and Bayesian subjective probability can assist me in pricing those bets. Moreover, inasmuch as there is any kind of precision to my uncertain knowledge and belief, I am more sympathetic to classical probabilism as a representation of that uncertainty than I am to other formal techniques in knowledge representation, for example the AGM axioms. And it seems to me that this system provides the most natural resolution of various problems related to partial belief, such as the preface paradox. Hence the three-tiered interpretation accords a place to Bayesian credences, defined in the standard way according to the Bayesian literature.

By contrast, I will use the term "Bayesian subjectivism" to denote the following expansion of the view:

1. An agent has at all times credences for all uncertain propositions, representing implicit dispositions to act, and forming a completed σ-algebra that is consistent according to the Kolmogorov axioms of probability.

2. All knowledge can be assimilated to this framework, and all learning can be described as conditional update.

I disagree intensely with this view.[5] As a frequentist, I am perpetually surprised by the insistence of Bayesian authors that I have credences for propositions I have never considered, that I should elevate my unfounded hunches and gut instincts to the level of formalized belief, or that I should apply the principle of indifference and believe completely uncertain propositions to degree $\frac{1}{2}$. When confronted with dispositional or gambling analyses, which allege that my credence can be measured by my propensity to bet,

[5]Binmore [2006], who has a similarly skeptical perspective, uses "Bayesian" to describe moderate views of the first type and "Bayesianite" for the second.

my response is that there are many propositions on which I would simply refuse to bet, or deny that I have a precise indifference price for a bet. And indeed, the rationality of this response is being defended increasingly in the literature, under the heading of "imprecise" or "mushy" credence — see Elga [2010] or Joyce [2010] for arguments that there are situations in which precise credences are unobtainable or unjustifiable.

Nor is the difficulty of eliciting precise credences the only foundational difficulty with the Bayesian view. The intuition that Bayesian subjectivism as an account of all uncertain reasoning represents an inherently unfeasible ideal, even at the aspirational level, is supported by both philosophical and mathematical evidence. Examples include Garber's observation [1983] that taking the position literally implies the existence of a unified language for all of science (the project the logical positivists failed to complete), or Paris's proof [1994] that testing probabilistic beliefs for consistency is NP-complete.

However, as in the case of physical chance, a variety of conceptions of Bayesian probability are enabled by the three-tiered interpretation. In particular, if you are a traditional Bayesian, then you have Bayesian credences for a very wide range of propositions. Some of your credences also happen to be frequency judgments, and some of those in turn happen to be chances, but these distinctions are not of central importance to you. But the three-tiered view also enables a much more skeptical attitude to Bayesian probability, one that is identifiable with the skepticism of traditional frequentism: credences that have frequency interpretations can take on definite values while credences that have no such interpretation are unsharp or remain in a state of suspended judgment.

7 The transfer principles

I claimed that probabilities from the first tier transfer to the second, and from the second to the third, the conjunction of these constituting a fragment of the Principal Principle. However, I suspect that no one will be especially interested in contesting this aspect of my argument — Bayesians already endorse the Principal Principle, and frequentists find it perfectly acceptable to bet according to frequency ratios. So the purpose of my discussion will be as much to clarify the underlying notions as to prove the principles.

Definition 3. *Let E be an event. If an agent can assign E to a reference class R, she knows a physical chance p for events in the class R, and she has no other relevant information, she is obliged to have a frequency judgment of p for E and R.*

This is more or less trivial. If we know that a class of events exhibits chance, then we can model sequences of those events as I.I.D. draws from

the relevant distribution.

Definition 4. *Let E be an event. If an agent has a frequency judgment p for E (by virtue of associating it with an unambiguous reference class R), and no other relevant information, he is obliged to have a Bayesian subjective probability of p for E.*

The argument for this is as follows: let the agent consider how to buy and sell bets for a sequence of n sequence of events in the reference class R, for n arbitrarily large. He estimates that a proportion p of these events will come true. Therefore, the fair price for the sequence of bets is pn; any higher and if he buys the bets at that price, he will lose money according to his estimate, any lower and he will lose money by selling them. But since R is epistemically homogeneous for the agent, and in particular he has no information that distinguishes E from the other events, each individual bet must have the same price. Thus, his fair price for a bet on E is $\frac{pn}{n} = p$. □

How much of the Principal Principle have we recovered? We have it for any event that has a chance and belongs to a reference class. This captures most conventional uses of PP, for example the radioactive decay of atoms (even Unobtainium). But we have seemingly failed to recover it in the case of one-off events. For example, what we have come to understand an inherently unique macrophysical phenomenon as possessing a chance of p? We cannot have a frequency judgment about it, so on the basis of the reasoning here we are not constrained to have a credence of p in it.

This is a genuine problem and I cannot resolve it entirely here — a solution would seemingly require a detailed analysis of the meaning of chance. As a last resort I can simply defer to an existing justification of PP that doesn't go through frequency judgments. But here is a brief sketch in defense of the full PP on the basis of the frequency judgments view. PP is inherently a principle of epistemology, not metaphysics, because it describes a constraint on credences (which are necessarily an epistemic notion). Therefore it is appropriate to ask how we would actually come to know the value of this one-off macrophysical chance — we couldn't have learned it from observed frequency data. The most natural answer seems to be that we would learn it via a theoretical model in which the overall macrophysical chance supervened on microphysical chances. And then this model would provide the basis for a frequency interpretation of the chance: over the reference class of situations satisfying the initial conditions of the model, the desired event would come to pass in some proportion p of the situations. This doesn't exhaust all possible methods by which we could come to know p, but I hope it fills in a good portion of the gap.

Finally, notice the qualifications in the second principle: the reference class must be unambiguous, and there must be no other relevant information. The second of these requirements corresponds to the requirement of

admissibility commonly associated with the Principal Principle; if you have information about an individual event that informs you about it beyond the background chance of success or failure, then PP is not applicable. (A simple example: you are playing poker and your opponent is trying to complete a flush. You know that the objective chance of this occurring is low, but you have seen him exhibit a "tell", for example, the widening of the eyes in excitement. Your credence that he has a flush should increase to a value higher than that dictated by the PP.) There is a sophisticated literature on when exactly PP is admissible, and I have no particular stance on the issue. Indeed, my view is that both qualifications are features and not bugs. When admissibility is debatable or the reference class is ambiguous, there is no fact of the matter about what should be believed.

8 Populations, direct inference, and the Principle of Indifference

White [2009] calls the second transfer principle "Frequency-Credence". He claims that it implies the generalized Principle of Indifference, i.e., the rule that if you are faced with n mutually exclusive alternatives and have no information to distinguish them, you should assume a credence of $\frac{1}{n}$ for each one. An especially revealing case is an individual proposition q concerning which you have no relevant information: since exactly one of $\{q, \neg q\}$ is true, the Principle of Indifference indicates that you should assign $P(q) = P(\neg q) = 0.5$. Such a principle is of course anathema to frequentists, since it is applicable in cases when there is no possible frequency interpretation of $P(q)$. Thus, White's purpose is to show that frequentist squeamishness about the Principle of Indifference is incoherent. Here is his statement of Frequency-Credence:

Definition 5. *If (i) I know that a is an F, (ii) I know that freq*$(G \mid F) = x$ *(the proportion of Fs that are G), and (iii) I have no further evidence bearing on whether a is a G, then* $P(a \text{ is a } G) = x$.

and here is his proof ($\simeq$ denoting epistemic indistinguishability):

> Let $F = \{p_1, p_2, \ldots, p_n\}$ be any set of disjoint and exhaustive possibilities such that $p_1 \simeq p_2 \simeq \ldots p_n$. Let G be the set of *true* propositions. For any p_i, (i) I know that p_i is an F; (ii) I know that *freq*$(G \mid F) = \frac{1}{n}$ (exactly one member of the partition $\{p_1, p_2, \ldots p_n\}$ is true); and (iii) I have no further evidence bearing on whether p_i is G (I am ignorant concerning the p_i, with no reason to suppose that one is true rather than another). Hence by FC, $P(p_i \text{ is a } G) = \frac{1}{n}$, i.e., $P(p_i \text{ is true}) = \frac{1}{n}$, so $P(p_i) = \frac{1}{n}$. $\square$

White challenges opponents of the Principle of Indifference to identify a restriction of Frequency-Credence that disallows this proof. Fortunately, the frequency judgments interpretation and the second transfer principle qualify as just such a restriction. Moreover, the precise way in which they block the conclusion reveals some interesting information.

Everything hangs on the following assertion in the proof: that $freq(G \mid F) = \frac{1}{n}$. For White, this is just the observation that exactly one of the possibilities $p_1 \ldots p_n$ is true, i.e., it is the finite frequency of true propositions among the available possibilities. But for the second transfer principle to apply, this must constitute a genuine frequency judgment, and without a reference class and a conception of repeated trial, a frequency judgment cannot exist. In particular, if the alternatives are q and $\neg q$ for a single-case proposition q with no obvious notion of trial ("God exists", "Chilperic I of France reigned before Charibert I"), no frequency judgment will be supported, and there is no obligation to set $P(q) = P(\neg q) = 0.5$; rather it is perfectly reasonable to be in a state of suspended judgment, or to have an unsharp credence interval.

There is a subtlety here because the principle of indifference can indeed be a source of legitimate frequency judgments. If for some genuine reference class of repeated trials, each trial has the same n mutually exclusive outcomes, then it can be perfectly legitimate to estimate a priori the long-run frequency of each one as $\frac{1}{n}$.[6] This estimation may not be justified or accurate, but that doesn't matter; as discussed previously, what matters is the possibility of confirming or disconfirming the judgment from empirical data. But even in this case, we do not recover the principle of indifference as an *obligation*, merely as an option. There is no obligation to formulate frequency judgments in the absence of evidence — dispositional betting arguments try to elicit credences in this way, but obviously this doesn't go through for frequency judgments.

8.1 Direct inference

There is another subtlety: observed finite frequencies are not necessarily frequency judgments! Consider the following scenario, discussed by Levi [1977] and Kyburg [1977]: of the 8.3 million Swedes, 90% of them are Protestants. Petersen is a Swede. What should our credence be that Petersen is a Protestant? Intuitively, there seems to be a frequency probability that P(Petersen is a Protestant) = 0.9. Arguments of this form — going from relative frequency in a population to a credence — are called *direct inferences* or *statistical syllogisms*, and they are a significant aspect of our probabilistic reasoning. But

[6] In a Bayesian framework, this would be an uninformative prior, or indifference prior, over the n alternatives.

if we try to phrase this as a frequency judgment, we encounter problems. The Swedes are not a reference class of events, and there is no obvious notion of repeated trial at work.

The situation seems analogous to the case of $\{q, \neg q\}$. The intuition that we should have a credence of 0.9 is seemingly grounded in the idea that Petersen is one of the 8.3 million Swedes, and we are indifferent as to which one he is. But if we allow unrestricted reasoning of this kind, then it will apply to the two propositions $\{q, \neg q\}$ as well, and White's challenge will succeed after all — we will have conceded that making use of frequency probabilities implies a generalized principle of indifference. Can we save the intuition that $P(\text{Petersen is a Protestant}) = 0.9$ without conceding $P(q) = P(\neg q) = 0.5$?

Here is a case that may clarify what the frequency judgments interpretation says about this kind of reasoning. You are a contestant on a game show; a prize is behind exactly one of three closed doors, and you must choose which one to open. What should your credence be that the prize is behind the left door? Whatever this credence is, if it is to be associated with a frequency judgment, it must be possible to clarify it with respect to the long-run behavior of repeated trials. The natural conception of repeated trial here is that we would play the game repeatedly and measure the proportion of times that the prize is behind the left door. And it is not clear that any particular frequency judgment is supported about this reference class of trials — we might imagine that the show host has a bias towards one of the doors in particular. Considerations like this support a view in which your credence that the prize is behind the left door is indeterminate or unsharp, or in which you suspend judgment about the question. Contrast this with the following claim: if you flip a fair coin with three sides and use the result to decide which door to choose, you have a frequency judgment of $\frac{1}{3}$ that this procedure will yield the correct door, regardless of what the host does. In this case, a frequency judgment is fully supported, because the reference class is clear (flips of the coin) and its properties are unambiguous, and there is a convincing case that the second transfer principle obligates you to have a credence of $\frac{1}{3}$.[7]

However, it seems natural that we should wish the credence of $\frac{1}{3}$ to be available at least as an *option* for the rational agent, and to be able to make sense of this under the frequency judgments interpretation. I think this is possible via the following expedient: we construct a model of the show in which the host selects the prize door via a coin flip. Acknowledging that this model, like any model, may not be true, we can use it to support a frequency judgment of $\frac{1}{3}$ for each door. Returning to our original problem, we can adopt a model in which the process by which we encounter Swedes

[7]This distinction is closely related to the Ellsberg paradox, and to the decision-theoretic and economic notions of ambiguity aversion and Knightian uncertainty.

is a chance process, analogous to a lottery in which we are equally likely to draw any individual Swede. This model then supports a frequency judgment of .9 for Protestants and a credence of .9 that Petersen is one.

This technique — modeling unknown processes as chance processes — is the general idea of how direct inference is supported under the frequency judgments interpretation. Does it, as White alleges, imply a generalized principle of indifference? As discussed above, even when the technique is applicable, it is not obligatory; the option of suspending judgment (or having an unsharp credence) is left open. Moreover, the technique seems to get at an important distinction between two kinds of indifference. It applies straightforwardly to situations where one is indifferent between *individuals* (prize doors, Swedes), but not to situations where one is indifferent between *propositions* (which king reigned first). Indeed, to interpret the second kind of indifference within our framework, we would seemingly have to talk about an indifference between *possible worlds*, and of a chance process deciding which one we live in. At this point we have regressed to the kind of reasoning decried by C. S. Peirce, of imagining that "universes [are] as plenty as blackberries" and we can "put a quantity of them in a bag, shake well them up, [and] draw out a sample." This kind of reasoning is not frequentist and therefore it is appropriate that we cannot understand it on frequentist terms.

9 Hájek's objections to frequentism

As I understand the frequency judgments interpretation, it avoids the bulk of Hájek's objections simply by failing to be a frequentism in the classical sense of the term. Let F.n denote his nth objection to finite frequentism, and H.n his nth objection to hypothetical frequentism. It seems to me that most of his objections are straightforwardly dismissed by one or more of the following concessions:

1. Not constraining frequency probabilities to be actual finite frequencies. This obviates objections F.2, F.5, F.6, F.8, and F.12-15.

2. Not considering frequency probabilities to be determined by hypothetical infinite sequences of trials. This obviates objections H.1-6, H.8-9, and H.13-14.

3. Acknowledging the possible existence of physical chance. This answers objections F.3, F.7, F.9, F.11, H.7, and H.10,

4. Acknowledging the legitimacy of Bayesian subjective probabilities. This answers objections F.10 and H.10.

Of the remaining objections: Hájek himself has subsequently repudiated F.1, which criticizes finite frequentism on the grounds that it admits a reference class problem. As discussed in section 5.2, Hájek now considers the reference class problem to affect every interpretation of probability, and I fully concur. I take H.11 (which concerns paradoxes associated with uncountable event spaces) to affect the Kolmogorov formalization of probability itself rather than frequentism specifically. H.15, which says that frequency interpretations cannot make sense of infinitesimal probabilities, I take to be a feature and not a bug.

The two remaining objections, F.4 and H.12, have a common theme — they say that frequentism cannot make sense of propensity probabilities. This is a serious issue that the three-tiered interpretation does not entirely address. In particular, here is Hájek's thought experiment from H.12:

> Consider a man repeatedly throwing darts at a dartboard, who can either hit or miss the bull's eye. As he practices, he gets better; his probability of a hit increases: P(hit on $(n+1)$th trial) $>$ P(hit on nth trial). Hence, the trials are not identically distributed. [...] And he remembers his successes and is fairly good at repeating them immediately afterwards: P(hit on $(n+1)$th trial | hit on nth trial) $>$ P(hit on $(n+1)$th trial). Hence, the trials are not independent.

Intuitively, all of these probability statements are meaningful, objective statements about the properties of the man (or of the dart-throwing process as a whole). Yet by their nature, we have difficulty in understanding them as statements about relative frequencies over sequences of independent and identically distributed trials. Hájek is unimpressed with the reply that in order to obtain a frequency interpretation of these probabilities, we should "freeze the dart-thrower's skill level before a given throw" and then consider hypothetical repeated throws by the frozen player. On one level, this notion of "freezing" involves an appeal to a nonphysical counterfactual. On another, relative frequencies seem irrelevant to the intuition that the thrower has, before each throw, some single-case *propensity* to hit or miss the target. The intuition here is analogous to the case of chance, except that there is no clear way to interpret the dart-throwing system as subject to physical chance.

I can see no way for the three-tiered interpretation other than to resolutely bite this bullet. That is to say, the three-tiered interpretation does not make rigorous the idea of propensity probabilities that are not chances. In this I am agreeing with von Mises, who held that we cannot make sense of such single-case assertions as "the probability that John will die in the next 5 years is 10%." In defense of this refusal with respect to Hájek's dart-thrower, I can only say this: the only way we were able to formulate this model in the first place was to observe the behavior of multiple dart-throwers, and thus

to reason about reference classes of darts players in specific situations (e.g., immediately after hitting the bulls-eye). Furthermore, how would we confirm the applicability of this model to any specific player? It seems that we would do so via some sort of calibration test — and, as discussed in section 5.5, calibration is always implicitly or explicitly dependent on some notion of frequency probability.

10 Advantages of the tiered interpretation

10.1 Statistical pragmatism

Two things are needed for the tiered interpretation to serve as a foundation for statistical pragmatism [Senn, 2011], i.e., for a worldview in which frequentist and Bayesian methods coexist. In order to admit the use of Bayesian methods, it must acknowledge the existence of non-frequentist prior probabilities. But it must also formally distinguish the probabilities used by properly frequentist methods from those used by Bayesian methods; otherwise, a frequentist method is simply a peculiarly defective Bayesian method. Thus, there are two claims: firstly, that the probabilities referred to in frequentist statistical methods are frequency judgments, and secondly, that Bayesian statistical methods are characterized by their use of probabilities that cannot necessarily be interpreted as frequency judgments.

Although I am not an expert in statistics, I am confident in both of these claims. A canonical example is classical significance testing. A p-value of .05 means we estimate that if the null hypothesis were true and we repeated the experiment, .05 of the experiments would exhibit results as extreme as the one observed. This is straightforwardly a frequency judgment. Other methods utilizing test statistics, such as chi-squared testing, follow this pattern; the test statistic is computed from the data and then an estimate is given for the proportion of experiments (given the null hypothesis) that would exhibit correspondingly extreme values of the statistic. With confidence intervals, the frequency judgment attaches to the procedure of deriving the interval: a 90% confidence interval is associated with the estimate that if we repeatedly sampled and computed the confidence interval, 90% of the resulting intervals would contain the true value of the parameter.

In contrast, Bayesian methods in general allow the use of probabilities that have no frequency interpretation. For example, a prior probability for a hypothesis will not have one in general; rather it will represent epistemic uncertainty about the truth of the hypothesis. Of course, there are settings in which the prior in a Bayesian method may be interpretable as a frequency judgment. Consider someone with three coins, with biases $\frac{1}{4}, \frac{1}{2}, \frac{3}{4}$, who draws one of them at random from an urn, flips it 10 times, and observes 6 heads.

The agent can begin with a uniform prior distribution that assigns probability $\frac{1}{3}$ to each coin, then use Bayes' rule to obtain posterior probabilities as to which coin he has. In this case, his prior is in fact a frequency judgment ("over a long sequence of urn drawings, each coin will be drawn $\frac{1}{3}$ of the time"), and thus his posteriors are also frequency judgments ("over a long sequence of drawing coins from the urn and flipping them ten times, of the times I see 6 heads, ≈ 0.558 of them will be because I drew the $\frac{1}{2}$-coin"). But the method would be equally applicable if the prior reflected only the agent's subjective degrees of belief in which coin he had.

10.2 Cromwell's rule

A notorious problem for subjectivist Bayesianism is the difficulty associated with assigning probabilities of 0 or 1. Let's say you assign $P(A) = 0$. Then for any B, $P(A \mid B) = \frac{P(A \cap B)}{P(B)} \leq \frac{P(A)}{P(B)} = \frac{0}{P(B)} = 0$, so you can never revise $P(A)$ by conditioning on new information. The case for $P(A) = 1$ is analogous, as is the situation when standard conditionalization is replaced by Jeffrey conditionalization.

Thus, according to many interpretations, a strict Bayesian should never assign a probability of 0 to an event, no matter how unlikely; Lindley calls this requirement Cromwell's rule. But frequency judgments are not affected by this problem, because they can be revised arbitrarily. Perhaps the clearest example is the case of estimating the bias of a coin, where we admit a third event besides heads and tails: it is physically possible that the coin might come to rest on its edge, or that the outcome of the flip might remain undetermined in some other way. A strict Bayesian is apparently committed to having prior probabilities for all of these events — and fixing $P(H) = P(T) = 0.5$ entails a violation of Cromwell's rule, since no probability mass is left over for them.[8] But under the frequency judgments interpretation, there is no difficulty associated with revising a probability from zero to a nonzero value.

Perhaps questions of this kind are artificial, unrelated to genuine concerns of statistical practice? On the contrary, they seem to correspond to actual methodological difficulties that arise when adopting a strictly Bayesian perspective. Gelman and Shalizi [2012] describe how a rigidly Bayesian outlook can be harmful in statistical practice. Since "fundamentally, the Bayesian agent is limited by the fact that its beliefs always remain within the support of its prior [i.e., the hypotheses to which the prior assigns nonzero probability]", it is difficult to make sense of processes like model checking or model

[8] One standard technique for dealing with this is to leave a small amount of mass over for a "catch-all" hypothesis, which is a disjunction over all seen and unseen alternate hypotheses. See Fitelson and Thomason [2008] for an argument that this is false to scientific practice.

revision, in which a model can be judged inadequate on its own merits, even before a suitable replacement has been found. They instead join Box and others in advocating a picture where individual Bayesian models are subjected to a non-Bayesian process of validation and revision. Dawid, whose calibration theorem suggests a similar difficulty with the Bayesian agent being able to recognize his or her own fallibility, is led also to an endorsement of Box. The point is not that these statisticians are betraying Bayesianism, it is that their pragmatic interpretation of Bayesian statistical methodology bears little resemblance to the worldview of the formal epistemologist who endorses Bayesian confirmation theory.

10.3 Foundations of conditional probability

Bayesian probability proves its worth in dissolving paradoxes associated with partial belief. Yet it is affected by its own set of paradoxes. I believe that the tiered interpretation, in its capacity as a relaxation of strict Bayesian discipline, can dissolve some of these as well — most notably, those in which Bayesian conditionalization is expected to subsume all probabilistic model-building.

Hájek [2007] gives the following paradox. An urn has 90 red balls and 10 white balls. Intuitively, P(Joe draws a white ball from the urn | Joe draws a ball from the urn) = .1. But in the standard Kolmogorov interpretation of probability, conditional probability is not a primitive notion but a derived notion, so in order for this statement to be true, we must have P(Joe draws a ball and it is white) / P(Joe draws a ball) = .1. But neither one of these unconditional probabilities appears well-defined on the basis of our assumptions. As Hájek asks, "Who is Joe anyway?"

Hájek's solution is to suggest that conditional probability is the true primitive notion and that we should consider alternate (non-Kolmogorov) formulations of probability that elevate it to its rightful place as such. But this seems to miss the mark. In particular, even though P(Joe draws a white ball | Joe draws a ball) is well-defined, P(Joe draws a white ball | Bill flips a coin) is not. Moreover, we can recover unconditional probability from conditional probability, for example by conditioning on independent events (e.g., P(Joe draws a white ball | a distant radium atom decays)) or on tautologies (e.g., P(Joe draws a white ball | $p \vee \neg p$)). It seems that conditionalization is orthogonal to the true problem: when does a situation support a probabilistic analysis?

Under the tiered interpretation of probability, this problem is confronted directly and admits a natural resolution. The fact that Joe is drawing a ball from the urn provides enough information to support a model and a frequency judgement: it calls into existence a probabilistic model in which we have an extremely simple event space: "Joe draws a white ball" or "Joe

draws a red ball". In this model, the value from our intuition appears as an unconditional probability: P(Joe draws a white ball) = .1. Saying this is no more and no less than saying that if Joe repeatedly draws balls from the urn with replacement, the natural estimate of the proportion of white balls is .1. In general, the process of assigning an event E to a reference class and then identifying $P(E)$ with the frequency judgment for that class is a more natural description of our probabilistic model-building than a strict Bayesian conditioning view.

Hájek's other paradox in the article, that of conditioning on events of probability zero, admits a similar resolution. Hájek has us consider a random variable X uniformly distributed on $[0, 1]$. Intuitively, $P(X = \frac{1}{4} \mid X = \frac{1}{4} \vee X = \frac{3}{4})$ equals $\frac{1}{2}$. But if we expand this using the standard definition of conditional probability, we get $\frac{P(X=\frac{1}{4})}{P(X=\frac{1}{4} \vee X=\frac{3}{4})} = \frac{0}{0}$, which is undefined.

Once again, the problem seems to be that we are taking an unnecessarily narrow view of the model-building process. It is natural that we should try to transform a continuous distribution into a discrete one by setting $P(X = a) = f(a)$, where f is the density function, and renormalizing — this has a natural interpretation as the outcome of considering $P(|X - a| < \epsilon)$ for smaller and smaller values of ϵ. When applied to Hájek's uniform distribution, with a ranging over $\{\frac{1}{4}, \frac{1}{2}\}$, this yields the expected answer $P(X = \frac{1}{4}) = P(X = \frac{3}{4}) = \frac{1}{2}$. It should not be considered problematic that this model transformation cannot be interpreted as a conditional update.

10.4 Sleeping Beauty

The Sleeping Beauty Paradox, popularized by Elga [2000], goes as follows. A fair coin, i.e., one that lands heads with an objective probability of $\frac{1}{2}$, is flipped on Sunday, and then Beauty is put to sleep. If it lands heads, Beauty is awakened on Monday, interviewed, his memory is wiped and he is put back to sleep. If it lands tails, this is done once on Monday and once on Tuesday. Beauty has just awoken. What should his credence be that the coin landed heads? The "halfer" position is that since the coin is fair, $P(H)$ must equal $\frac{1}{2}$. But if the experiment is repeated many times, only $\frac{1}{3}$ of Beauty's awakenings will be because the coin landed heads — hence the "thirder" position that $P(H) = \frac{1}{3}$. Which of these is the correct credence?

Sleeping Beauty is a vexing problem for Bayesian epistemologists and has generated a rich literature. But, as Halpern [2004] observed, the paradox is immediately dissolved by a frequentist analysis: it is a pure instance of reference class ambiguity. If Beauty analyzes his situation using the reference class of all coinflips, then the probability of a head is $\frac{1}{2}$. If he analyzes it instead using the reference class of all awakenings, the probability of a head is $\frac{1}{3}$. Under the tiered interpretation, there are thus two possible frequency

judgments, one with value $\frac{1}{2}$ and one with value $\frac{1}{3}$. But since the reference class is ambiguous, neither one passes down to become a credence. For a frequentist (or anyone who is free to suspend judgment about credences), the problem is simply one of vagueness.

This seems unsatisfying. After the frequentist throws up his hands in this way, how should he bet? As Halpern shows, the fact is that there exist Dutch Books against both "halver" and "thirder" agents, but they are not true Dutch Books: they rely on the ability of the bookie to vary the number of bets that are bought and sold according to the number of awakenings. Therefore the ideal betting behavior is not fixed, but depends on the capabilities of the adversary.

Beauty has genuine probabilistic knowledge about his situation: over the long run, half of all fair coin tosses are heads, and a third of his awakenings are because the coin landed heads. And he can, in fact, use this knowledge to buy and sell bets on H. For example, Beauty can buy and sell bets on heads on Sunday, and the fair price for those bets will be $\frac{1}{2}$. And if Beauty has an assurance that the exact same bets on heads will be on offer every time he wakes up (perhaps they are sold from a tamper-proof vending machine in the laboratory), the fair price for those bets will be $\frac{1}{3}$. What Beauty cannot safely do is fix a single indifference price and then buy and sell bets at that price, i.e., act in accordance with the traditional operational definition of credence. Beauty can have probabilistic knowledge about H without having a credence.[9]

10.5 White's coin puzzle

White [2009] is committed to the Principle of Indifference, in particular as an alternative to the suspension of judgment about credences. His thought experiment of the "coin puzzle" is intended to show that suspension of judgment is unsatisfactory. As with the previous discussion of White in section 8, the onus is on the frequentist to reply.

> You haven't a clue as to whether q. But you know that I know whether q. I agree to write "q" on one side of a fair coin, and "$\neg q$" on the other, *with whichever one is true going on the heads side* (I paint over the coin so that you can't see which sides are heads and tails). We toss the coin and observe that it happens to land on "q".

[9] I believe that Beauty will be protected against a variety of adversaries by having an unsharp credence interval of $[\frac{1}{3}, \frac{1}{2}]$, but formulating and proving this is beyond the scope of this paper.

Let P denote your credence function before seeing the flip, and P' your credence function afterwards. Let H denote the event that the coin lands heads. White notes that the following statements are jointly inconsistent:

1. $P(q)$ is indeterminate, i.e., before seeing the flip, you have no precise credence that q. (One natural formalization of this is to say that $P(q)$ is interval-valued, e.g., $P(q) = [0, 1]$. This can be read as "my credence in q is somewhere between 0 and 1.")

2. $P(H) = \frac{1}{2}$, i.e., before seeing the flip, you have a precise credence of $\frac{1}{2}$ that the coin will land heads.

3. $P'(q) = P'(H)$. This should be true because after seeing the flip, q is true if and only if the coin landed heads.

4. $P(q) = P'(q)$. This should be true because seeing the flip provided no information about whether q is in fact true. (Note that this would be false for a biased coin.)

5. $P(H) = P'(H)$. This should be true because seeing the flip provided no information about whether the coin landed heads. (Note that this would be false if you had meaningful information about p, in particular a a sharp credence of anything other than $\frac{1}{2}$.)

Put these together and we derive $P(q) = P'(q) = P'(H) = P(H) = \frac{1}{2}$, contradicting claim 1. White's conclusion is to deny that 1 is rationally permissible — rather, we should begin with a sharp credence of $P(q) = \frac{1}{2}$ via the Principle of Indifference. What should the proponent of unsharp credences do instead? Joyce [2010] moves instead to deny claim 5 and set $P'(H)$ to equal $P(q)$. Paradoxically, this causes an *dilation* of your credence in $P(H)$ — your $P(H)$ was precisely $\frac{1}{2}$ but your $P'(H)$ has become unsharp or interval-valued. Seeing the coin land has apparently reduced your knowledge!

My response to the coin puzzle is to affirm Joyce's view and accept dilation, combined with the rule (maximin expected utility) given by Gärdenfors and Sahlin [1982] for betting on unsharp credences. According to this view, the correct action for an unsharp agent with credence interval $P(q) = [0, 1]$ is as follows: before seeing the outcome of the flip, it is permissible to buy and sell bets on H for 0.5, to buy bets on p for prices ≤ 0, and to sell bets on p for prices ≥ 1. After the outcome of the flip has been revealed, your betting behavior for H should dilate to match your behavior for p. But, on my view, it is only your credences that dilate — your frequency judgment that $P(H) = \frac{1}{2}$ is exactly the fragment of your knowledge that is not destroyed by seeing the p-side of the coin come up.

This is the "conservative betting" behavior that White discusses and rejects. His argument against it uses a scenario of long-run betting on repeated

instances of the coin puzzle, with a series of coin flips $heads_i$ and a different unknown proposition p_i each time:

> On each toss you are offered a bet at 1:2 [i.e., for a price of $\frac{1}{3}$] on $heads_i$ once you see the coin land p_i or $\neg p_i$. Since your credence in $heads_i$ is mushy at this point you turn down all such bets. Meanwhile Sarah is looking on but makes a point of covering her eyes when the coin is tossed. Since she doesn't learn whether the coin landed p_i her credence in $heads_i$ remains sharply $\frac{1}{2}$ and so takes every bet [....] Sure enough, she makes a killing.

This hinges on an ambiguity in how exactly the bets are being offered. If you know for certain that the bets will be offered, i.e., if you have a *commitment* from your bookmaker to sell the bets, then that is equivalent to the bets being offered before the coin is tossed, and you are justified in buying them. But if your bookmaker can choose whether or not to offer the bet each time, you would be very ill-advised to buy them, since he can then offer them exactly in the cases when he knows that $\neg p_i$, and you will lose your $\$\frac{1}{3}$ every time! This is exactly the situation that unsharp credences are intended to prevent: if you suspend judgment and refuse to bet, you can't be taken advantage of. And once the p_i or $\neg p_i$ side of the coin has been revealed, you can be taken advantage of by someone who knows the truth about p_i, so you should stop buying and selling bets.[10] But what has changed is your betting behavior about H, not your knowledge about H. Your knowledge is exactly your frequency judgment and it remains intact.

The coin puzzle is a powerful illustration of the following fact: for an unsharp agent, probabilistic knowledge and betting behavior can come apart. Thus, it is only paradoxical under behaviorist interpretations of probability in which they are held to be synonymous. My hope is that the three-tiered interpretation distinguishes the two in a natural way, and in a way that affirms the core intuitions of frequentism.

11 Acknowledgements

I am grateful to Sherri Roush, Alan Hájek, Roy Frostig, Jacob Steinhardt, and Jason Auerbach for helpful discussions.

[10]There is, however, no need to revoke or cancel any existing bets, as White alleges in a subsequent thought experiment.

References

Ken Binmore. Making decisions in large worlds. URL `http://else.econ.ucl.ac.uk/papers/uploaded/266.pdf`. 2006.

Rudolf Carnap. The two concepts of probability: The problem of probability. *Philosophy and Phenomenological Research*, 5(4):513–532, 1945.

A. Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Antony Eagle. Deterministic chance. *Noûs*, 45(2):269–299, 2011.

Adam Elga. Self-locating belief and the sleeping beauty problem. *Analysis*, 60(2):143–8211, 2000.

Adam Elga. Subjective probabilities should be sharp. *Philosopher's Imprint*, 10(5), 2010. URL `http://www.princeton.edu/~adame/papers/sharp/elga-subjective-probabilities-should-be-sharp.pdf`.

Branden Fitelson and Neil Thomason. Bayesians sometimes cannot ignore even very implausible theories (even ones that have not yet been thought of). *Australasian Journal of Logic*, 6:25–36, 2008.

Daniel Garber. Old evidence and logical omniscience in Bayesian confirmation theory. In *Testing Scientific Theories*, volume X of *Minnesota Studies in the Philosophy of Science*, pages 99–131. University of Minnesota Press, 1983.

Peter Gärdenfors and Nils-Eric Sahlin. Unreliable probabilities, risk taking, and decision making. *Synthese*, 53(3):361–386, 1982.

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics in the social sciences. In *The Oxford Handbook of Philosophy of Social Science*. Oxford University Press, 2012.

Luke Glynn. Deterministic chance. *British Journal for the Philosophy of Science*, 61(1):51–80, 2010.

Alan Hájek. The reference class problem is your problem too. *Synthese*, 156 (3):563–585, 2007.

Alan Hájek. "mises redux" — redux: Fifteen arguments against finite frequentism. *Erkenntnis*, 45(2-3):209–27, 1996.

Alan Hájek. Fifteen arguments against hypothetical frequentism. *Erkenntnis*, 70(2):211–235, 2009.

Joseph Halpern. Sleeping beauty reconsidered: Conditioning and reflection in asynchronous systems. In *Oxford Studies in Epistemology*, volume 1, pages 111–142. Oxford University Press, 2004.

Carl Hoefer. The third way on objective probability: A sceptic's guide to objective chance. *Mind*, 116(463):549–596, 2007.

E. T. Jaynes. Some random observations. *Synthese*, 63(1):115–138, 1985.

James M. Joyce. A defense of imprecise credences in inference and decision making1. *Philosophical Perspectives*, 24(1):281–323, 2010.

Robert E. Kass. Statistical inference: The big picture. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 26(1):1, 2011.

Henry E. Kyburg. Randomness and the right reference class. *Journal of Philosophy*, 74(9):501–521, 1977.

Isaac Levi. Direct inference. *Journal of Philosophy*, 74(1):5–29, 1977.

David Lewis. Humean supervenience debugged. *Mind*, 103(412):473–490, 1994.

J. B. Paris. *The Uncertain Reasoner's Companion*. Cambridge University Press, Cambridge, UK, 1994.

Wesley C. Salmon. *Statistical Explanation & Statistical Relevance*. University of Pittsburgh Press, 1971.

Jonathan Schaffer. Deterministic chance? *British Journal for the Philosophy of Science*, 58(2):113–140, 2007.

Teddy Seidenfeld. Calibration, coherence, and scoring rules. *Philosophy of Science*, 52(2):274–294, 1985.

Stephen Senn. You may believe you are a bayesian but you are probably wrong. *Rationality, Markets and Morals*, 2(42), 2011.

Bas van Fraassen. Calibration: A frequency justification for personal probability. In *Physics, Philosophy, and Psychoanalysis*. D. Reidel, 1983.

Roger White. Evidential symmetry and mushy credence. In *Oxford Studies in Epistemology*. Oxford University Press, 2009.