Proving the Principal Principle

Wolfgang Schwarz[*]

01 August 2012

1 The Challenge

A noteworthy feature of objective probability, or chance, is its connection to rational belief. If you know that the coin you're about to toss has a 50% chance of landing heads, then you should give the same degree of belief to heads and tails. In general, objective probability satisfies the following *Coordination condition*, where $Ch(A)\!=\!x$ is the proposition that the chance of A equals x.

> **Coordination (first pass)**
> Normally, if P is a rational prior credence function, A a proposition, and $P(Ch(A)\!=\!x) > 0$, then $P(A/Ch(A)\!=\!x) \approx x$.

This fact about prior credence often carries over to posterior credence because information about chance tends to screen off other information relevant to the outcome of a chance process. For example, the information that the previous toss of your coin came up heads should not affect your credence in the next outcome, if you also know that the chance of heads is 1/2. In general, objective chance satisfies the following *Resiliency condition* (compare [Skyrms 1980]).

> **Resiliency (first pass)**
> Normally, if P is a rational prior credence function, A a proposition, and $P(Ch(A)\!=\!x) > 0$, then there is a substantial range of propositions B such that $P(A/Ch(A)\!=\!x \wedge B) \approx P(A/Ch(A)\!=\!x)$.

Combining Coordination and Resiliency, we get a form of the *Principal Principle* (compare [Lewis 1980]).

> **Principal Principle (first pass)**
> Normally, if P is a rational prior credence function, A a proposition, and $P(Ch(A)\!=\!x) > 0$, then there is a substantial range of propositions B such that $P(A/Ch(A)\!=\!x \wedge B) \approx x$.

[*] Thanks to John Barker, Kenny Easwaran, Alan Hájek, Marcus Hutter, Michael Smithson and Alastair Wilson for helpful comments and discussion.

Lewis's formulation in [Lewis 1980] isn't restricted to "normal" cases, and it states a strict rather than approximate equality. However, on Humean accounts of chance, the strict Principle is arguably false for certain "undermining" propositions A. Lewis's answer was to move to a more complicated "New Principle" (see [Lewis 1994]). For reasons that will become clear, I prefer to stay with a softened version of the old Principle.

In [Lewis 1980], Lewis also suggests that (i) chance should be indexed to a time t and world w, and that (ii) resiliency holds for all propositions B about the history of w up to t. (These propositions Lewis calls *admissible*.) I agree that chance should be indexed, although not necessarily to a time and a world. This requires some adjustments to the three principles above, which will be made in section 3. I do not follow Lewis in specifying a fixed domain of resiliency, mostly because I want my principles to cover not only forward-looking dynamical probabilities in fundamental physics, but also probabilities found in genetics, population dynamics or statistical mechanics. As [Skyrms 1980: 10–19] points out, every statistical theory comes with its own domain of resiliency, so there is little more we can say in full generality except that the domain includes a substantial range of propositions, including many propositions that one can easily come to know and that would otherwise be relevant to the proposition at hand.

There is a debate over whether probabilities in genetics, population dynamics or statistical mechanics deserve the name 'chance'. I do not want to enter this debate here. Whatever we call them, these probabilities display essentially the same connection to rational belief as dynamical probabilities in fundamental physics.

This connection puts a tight constraint on the interpretation of objective probability. Most probability functions do not satisfy Coordination and Resiliency. Lewis [1994] argued that any proposal to identify chance with some objective measure X must explain why X satisfies the Principal Principle. He conjectured that only Humean interpretations like his "best system" account can live up to this challenge. However, Lewis never explained why best system probabilities satisfy the Principle. Every now and then, someone tries to fill the gap. [Mellor 1971] and [Howson and Urbach 1993] try to derive the Principle for hypothetical frequentism, [Loewer 2004] and [Hoefer 2007] for versions of the best system analysis, [Deutsch 1999] and [Wallace 2012] for branch amplitudes in Everettian quantum mechanics. But many remain unconvinced by these attempts. Indeed, there is a growing consensus that the task is impossible – that no account of chance, Humean or non-Humean, can explain the Principal Principle (see e.g. [Black 1998], [Strevens 1999], [Hall 2004]). In this note, I want to make another attempt at showing that Humean chances satisfy the Principal Principle.

Let me clear about the goal. The point of the exercise is not to justify the Principal Principle. The Principle may well be an analytic truth about chance and credence. It says is that chance plays a certain role. Now when we consider different hypotheses about the nature of chance, we have to ask whether these candidates are apt to play the role. By analogy, consider the claim that Jack the Ripper (if he exists) committed such-and-such murders in the late 19th century. This may well be analytic. Nonetheless, when we consider different hypotheses about the identity of Jack the Ripper – that he is Lewis Carroll, or Barack Obama, or some metaphysically primitive entity – we have to ask whether there is any reason to believe that these candidates satisfy the Jack the Ripper role, i.e. that they committed the relevant murders.

The goal, then, is to show that on a given interpretation of 'chance', the Principal Principle follows from independently plausibly assumptions about rational belief – assumptions that do not involve the concept of chance. As we will see, this is not too hard if chance is interpreted as the characteristic function of the set of truths, or as relative frequency; the relevant mathematical theorems are mostly well-known and will be briefly reviewed in section 2. In section 3, I suggest a generalisation of the Principal Principle that allows for indexed chance and does not require chance to deal with "single cases". Afterwards, I return to the task of deriving the (now generalised) Principle, using first a frequentist and then a best system analysis of chance.

2 First steps

Let's begin with a simple case: the *fatalist* interpretation of chance. According to fatalism, only what in fact will happen has any chance of happening, and its chance is 1. The chance function Ch is the "omniscient" function that maps every true proposition to 1 and every false proposition to 0. (The domain may be somewhat restricted, to avoid Liar-type paradoxes that arise if the "propositions" are sentences.) On this interpretation, $Ch(A) = 1 \leftrightarrow A$ and $Ch(A) = 0 \leftrightarrow \neg A$ are analytic, hence the Coordination condition $P(A \mathbin{/} Ch(A) = x) \approx x$ reduces to $P(A \mathbin{/} A) \approx 1$ and $P(A \mathbin{/} \neg A) \approx 0$. Given $P(A) > 0$ and $P(\neg A) > 0$, these are basic theorems of the probability calculus. Resiliency is also guaranteed, since $P(A \mathbin{/} A \wedge B) = 1$ and $P(A \mathbin{/} \neg A \wedge B) = 0$ are theorems for all B with $P(A \wedge B) > 0$ and $P(\neg A \wedge B) > 0$.

Fatalist chances are completely determined by the history of actual outcomes in the world. So here we have a Humean account that does entail the Principal Principle, without any additional assumptions about rational credence. The only problem is that fatalism is not a plausible interpretation of the probabilities in statistical theories. (Incidentally, this shows that *pace* [Lewis 1980], the Principle does not exhaust our concept of chance.)

On the other hand, the fatalist result points towards a more general lesson. The fatalist chance function is a maximally informed credence function: it is the credence of an imaginary agent who knows absolutely everything about the world. On the best system account, the chance function can also be understood as very well-informed credence, corresponding to the beliefs of an imaginary agent who has access to all occurrent facts, but limited memory, so that she cannot store all these facts one by one. If the Principal

Principle holds for maximally informed credence, does it also hold for lesser credence functions?

Here is a reason to think that it does. Let P be a rational prior credence function, and define Ch as P conditioned on the true answer to some question Q. For simplicity suppose the possible answers to Q form a finite partition. Let A be any proposition and x any number so that $P(Ch(A){=}x) > 0$. Let E_x be the disjunction of all possible answers E to Q with $P(A/E){=}x$. (There is at least one such E, as otherwise $Ch(A){=}x$ would be impossible and couldn't have probability > 0.) Assuming that the answers are mutually exclusive, it is not hard to show that $P(A/E_x) = x$. Moreover, $P(Ch(A){=}x \leftrightarrow E_x){=}1$. So $P(A/Ch(A){=}x) = P(A/E_x) = x$. Thus if Ch is a probability function that lies in between the rational prior credence P and the omniscient function in terms of what it knows about the events in a world, then Ch and P are linked like chance and rational credence, insofar as they satisfy Coordination.

This is encouraging, but it does not go far enough. For one thing, we also need to establish Resiliency. Moreover, most Humeans do not define chance as rational credence conditional on the answer to a certain question – although the "pragmatist" best system accounts of [Cohen and Callender 2009] and [Frigg and Hoefer 2010] come rather close.

The perhaps best known Humean account of chance is (finite) *frequentism*. Here chance is identified with relative frequency in a suitable sequence of events. For example, if 40% of the coin tosses in a certain sequence land heads, then the chance of heads, relative to that sequence, is 0.4. Despite various foundational problems and unintuitive consequences, the frequency interpretation is very popular among scientists, and until recently was the standard account of objective probability in philosophy.

Bruno de Finetti [1937] proved some important connections between rational belief and relative frequency. Consider a sequence of n coin tosses, with 2^n possible outcomes, represented by the vectors $\Omega = \{H, T\}^n$. Suppose the rational prior belief function P regards these outcomes as *exchangeable*, meaning that it assigns the same probability to any two outcomes $\omega, \omega' \in \Omega$ with the same number of heads and tails. Let $Ch(H) = r$ be the proposition that the relative frequency of heads in the sequence is r (for $r \in \{\frac{m}{n} : 0 \leq m \leq n\}$). Let H_i be the proposition that the ith outcome is heads (for $i \leq n$). Exchangeability then entails that

$$P(H_i/Ch(H){=}r) = r.$$

Moreover, the longer the sequence, the more stable $P(H_i/Ch(H){=}r)$ becomes under conditioning on other outcomes:

$$P(H_i/Ch(H){=}r) \approx P(H_i/Ch(H){=}r \wedge H_j) \approx P(H_i/Ch(H){=}r \wedge \neg H_j).$$

So exchangeability guarantees both Coordination and Resiliency, of a kind.

"Of a kind", because the chance function Ch and the credence function P apply to different objects. Credence is defined for token events, such as the first toss landing heads, H_1. Chance, on the other hand, only applies to event types, e.g. heads. Since there is only one first toss, it is not very useful to measure the relative frequency of the first toss landing heads.

The correct response to this "type conflict" between chance and credence, I think, is to reformulate the Principles of section 1. This will be the topic of section 3. For now, let's choose a simpler response and extend the frequentist interpretation to events like H_1, by ignoring the reference class problem and letting every token event inherit its chance from the corresponding type, in this case H. To get a probability distribution over the space of outcomes Ω, we also need to specify joint probabilities for different tosses. The most natural choice here is to treat them all as independent. Thus suppose r is the ratio of heads in the actual sequence. Then for any sequence $\omega \in \Omega$,

$$Ch(\omega) = r^h(1-r)^{n-h},$$

where h is the number of heads in ω.

In this way, every relative frequency r determines an "extended" frequentist chance function Ch_r over Ω. De Finetti showed that as n goes to infinity, any credence function P that regards the sequence as exchangeable converges to a mixture of such chance functions. In fact, P comes close such a mixture as long as n is not very small. Hence if the underlying sequence is sufficiently long and $Ch{=}Ch_r$ is the hypothesis that Ch_r is the extended frequentist chance function (i.e., that r is the relative frequency of heads), then for all $A \subseteq \Omega$, the prior credence in A equals the expectation of the chance of A:

$$P(A) \approx \sum_r Ch_r(A)P(Ch{=}Ch_r).$$

This is weaker than the Coordination principle $P(A/Ch(A) = x) \approx x$. The principle itself fails, because Ch_r generally assigns positive probability to outcomes in which the relative frequency of heads is not r. For example, if A is the proposition that all tosses land heads, then $Ch(A){=}x$ entails that the actual ratio of heads is $\sqrt[n]{x}$. For $x < 1$, this means that $Ch(A){=}x$ is incompatible with A, so that $P(A/Ch(A){=}x) = 0$.

This is an instance of the *undermining problem* for Humean accounts of chance. In response, Hall [1994] and Lewis [1994] argued that the Principal Principle should be replaced by a more accurate "New Principle". In the present context, the "New" form of Coordination could be expressed as

$$P(A/Ch{=}Ch_r) = Ch_r(A/Ch{=}Ch_r).$$

It is easy to see that this is entailed by exchangeability: $Ch{=}Ch_r$ is the set of outcome sequences ω in which the relative frequency is r; Ch_r is uniform over this set, and so is P if it satisfies exchangeability.

More interestingly, we can also establish restricted, *ceteris paribus* versions of the old Principle. In particular, we have

$$P(H_1 \dots H_k / Ch(H_1 \dots H_k) = x) \approx x$$

as long as k is small compared to n. And we have Resiliency in the sense that for $j \neq i$,

$$P(H_i/Ch(H_i){=}r) \approx P(H_i/Ch(H_i){=}r \wedge H_j) \approx P(H_i/Ch(H_i){=}r \wedge T_j).$$

We assumed that the rational prior credence P regards the relevant sequence as exchangeable. Is this plausible? To be sure, if you regard a series of events as chancy and independent (with respect to the chances), then your beliefs about the outcomes should treat them as exchangeable. But we must explain why this should be so *on the frequentist interpretation of chance*. In fact, the assumption cannot be true. For suppose you learn that the first 100 outcomes in some binary sequence about which you have no preconceptions are 101010 . . . 10. Intuitively, you should then be more confident that next two outcomes are 10 than that they are 01. But then 101010 . . . 1010 has higher prior probability than 101010 . . . 1001, and the sequence is not exchangeable. Relatedly, frequentist definitions of chance are typically restricted to *random* sequences, rather than regular patterns like 101010 . . . 10. The information that the frequentist chance of heads in a given sequence is 0.5 therefore entails that the sequence is sufficiently random. It follows that exchangeability fails, since 101010 . . . 10 is ruled out.

Fortunately, the exchangeability assumption can be weakened. The above results still go through if we only require that the rational prior credence gives equal probability to any two *sufficiently random* sequences with the same ratio of outcomes. This restricted form of exchangeability is quite plausible: if all you know about a sequence is that it looks random and contains a certain ratio of 1s to 0s, you should arguably assign equal credence to random-looking sequences with that ratio.

A precise frequentist analysis would now have to define a suitable notion of randomness. For the present topic, these details are not terribly important. However, it may be worth adopting a further requirement on suitable sequences, which arguably follows from von Mises's [1919] characterisation of randomness: the relative frequencies should not fluctuate much between earlier and later parts of a sequence. This means that the sequence of states in a Markov chain are often not a suitable base for the identification of chance with relative frequency. A suitable base would here be the sequence of state transitions rather than the sequence of states. (Exchangeability with respect to such "derived" sequences is closely related to de Finetti's notion of *partial exchangeability*; see [Diaconis and Freedman 1980]).

In practice, when we toss a coin, or set up a Stern-Gerlach experiment, we rarely identify the occasion as the ith member of a certain sequence. Even if there is a privileged way of arranging the relevant events in a series, we rarely know at which position in the series we are. This is not a big problem for the current results, because the probability that the *present* toss lands heads, given $Ch(H) = x$, is plausibly the average of the conditional probability that the ith toss lands heads, weighted by the probability of the present toss being the ith toss.

On the other hand, we can also exploit our ignorance of the present position directly to support the Principal Principle, without assuming any form of exchangeability. This was already pointed out by Bertrand Russell in [Russell 1948: 402ff.] – to my knowledge, the first explicit outline of a proof for the Principal Principle. Russell's argument goes as follows. Let C be a class of events, and D an arbitrary member of C. Given that r percent of Cs are H, what is your degree of belief that D is H? Answer: r. This is a consequence of the "arbitrariness" of the choice. Now when we toss a coin, we haven't literally chosen an arbitrary member of the relevant class – whatever that would mean. But our epistemic situation with respect to the outcome is typically just the same: if r percent of tosses land heads, and we have no further information about this particular toss, then it could just as well be any member of the class, so our credence in heads should be r.

I will return to this line of thought in sections 4 and 5, where I will also explain how the results established so far bear on the best system analysis. But first, I want to make the promised amendments to the Principal Principle.

3 Generalising the Principal Principle

As formulated in section 1, Coordination and Resiliency presuppose that the objects of chance are unrepeatable, single-case propositions like *heads on the 17th toss*, rather than repeatable event types like *heads*. As we've seen, this is incompatible with the most straightforward frequentist interpretation. It arguably also doesn't match the probability statements in actual statistical theories, which typically say that under such-and-such conditions C, outcome A has probability x. Here, C and A are naturally understood as event types – in other words, *properties* – that can be instantiated several times within a world. The statistical law that outcome A under condition C has probability x can be understood as a "partial" counterpart of the strict law that all Cs are As. Like the strict law, the statistical law primarily states a relation between properties.

The principles from section 1 also neglected the relational aspect of chance: they did not take into account that chance might be relative to a condition, a time, or a reference class. This is not what we find in many statistical theories, and it contradicts several important accounts of chance. For Lewis, chance is relative to a time-world pair or history. For frequentists, chance is relative to an underlying sequence. For propensity theorists, chance is a measure of the causal tendency of a physical system in state C to produce outcome A.

The chance of A relative to C is a kind of conditional probability, but it need not satisfy standard laws for conditional probabilities. Without committing to a particular theory of chance, all we can say is that chance is a family of probability functions, indexed by a set of properties. More precisely, a chance function Ch maps every element C of some set Γ of mutually exclusive properties to a probability space $\langle \Omega_C, \mathcal{F}_C, Ch_C \rangle$, where the Ω_C is a set of mutually exclusive properties, $\mathcal{F}_C$ is a suitable algebra over Ω_C, and Ch_C is a probability measure on $\mathcal{F}_C$. For instance, if C picks out a certain type of die toss, Ω_C might be the set of possible outcomes $\{\mathit{One}, \mathit{Two}, \ldots\}$, and $\mathcal{F}_C$ the set of subsets of Ω_C. Intuitively, the set $\{\mathit{Two}, \mathit{Four}, \mathit{Six}\}$ here stands for the unspecific property of landing with an even-numbered side up. I will often refer to the members of Γ as *conditions* and those of Ω_C as *(basic) outcomes*, but these names aren't meant to carry any significance: a condition may, for example, simply be a time.

The probability $Ch_C(A)$ of A relative to C is not derived from an unrelativised probability distribution on a more inclusive algebra, perhaps as the ratio $Ch(A \wedge C)/Ch(C)$. We often have a well defined probability $Ch_C(A)$ e.g. of future states given present states, but no converse probability $Ch_A(C)$ of present states given future states, nor an unrelativised probability of the present state C. Relative to each index $C \in \Gamma$, there are ordinary conditional probabilities. Thus in the die toss example, it might be that $Ch_C(\mathit{Two}/\mathit{Even}) = Ch_C(\mathit{Two} \wedge \mathit{Even})/Ch_C(\mathit{Even}) = 1/3$. This must not be confused with $Ch_{C \wedge \mathit{Even}}(\mathit{Two})$ or $Ch(\mathit{Two}/C \wedge \mathit{Even})$, both of which are undefined. Confusing the two kinds of conditionality leads to a problem known as "Humphreys' Paradox".

Traditional, single-case propositions are properties of a special kind: the proposition A is the property of being such that A. So the present proposal does not rule out irreducible single-case chance. For the sake of generality, we might also allow for unrelativised chance. In this case, Γ is best identified with $\{\top\}$, the singleton of the tautologous property $\top$. A has unrelativised chance x iff A has chance x under condition $\top$. The condition is logically guaranteed to always obtain, so the relativisation does no work.

Now return to Coordination, Resiliency and the Principal Principle. We somehow need to include the extra argument place for conditions. The most obvious way to adjust Coordination would be to replace $P(A/Ch(A){=}x) \approx x$ with

(1) $\quad P(A/Ch_C(A){=}x \wedge C) \approx x.$

This resembles the "reformulated" Principal Principle in [Lewis 1980: 97]. Lewis's principle says, in effect, that

(2) $\quad P(A/Ch{=}f \wedge C) = f_C(A),$

where $Ch = f$ is the proposition that the function f is the chance function, and the condition C is a complete history of a universe up to a certain time. [Meacham 2005] recommends generalising this Principle to other conditions, in order to accommodate chances in statistical mechanics and time-symmetrical versions of quantum mechanics. Like Lewis, Meacham assumes that the conditions Γ are very rich in information. (Meacham even assumes that each $C \in \Gamma$ uniquely determines the true chance function Ch.) (1) and (2) then guarantee a great deal of resiliency. To illustrate, suppose Γ is a finite set of hypotheses about the complete history of the world up to now. Let $Ch_{@}(A)\!=\!x$ be the proposition that the chance of A relative to the history of the actual world up to now is x. By the law of total probability, and the fact that $Ch_C(A)\!=\!x \wedge C$ is equivalent to $Ch_{@}(A)\!=\!x \wedge C$,

$$P(A/Ch_{@}(A)\!=\!x) = \sum_{C \in \Gamma} P(C)P(A/Ch_C(A)\!=\!x \wedge C).$$

By (1), it follows that $P(A/Ch_C(A)\!=\!x \wedge C) \approx x$ for all C. Hence if E is any disjunction of propositions in Γ – in other words, any information about the past – then $P(A/Ch_{@}(A)= x \wedge E) \approx x$. So information about chance screens off all information about the past.

It would be nice, I suppose, if we could always identify a chance function's domain of resiliency with the relevant conditions Γ. However, I do not want to assume that information about chance always screens off information about the relevant condition. I also don't want to assume that the conditions in Γ are informationally rich. If a statistical theory specifies probabilities for outcomes of coin tosses, then the relevant condition is *being a coin toss*, or perhaps *being a coin toss of type so-and-so*. The instantiation of this condition entails very little about the world, so Coordination does not automatically entail Resiliency.

In any case, there is something odd about (1) and (2). If A is *landing heads* and C *being a coin toss*, then (1) constrains a rational agent's prior degree of belief in *landing heads*, conditional on *being a coin toss*. But what does it mean to assign degrees of belief to a property?

[Lewis 1979] argued that the objects of credence really are properties. On this view, to assign high credence to a property is (roughly speaking) to self-attribute that property. But this is not what we want. Most of us are fairly certain that we are not coin tosses, so it is not very useful to learn what we should believe conditional on this outlandish assumption. In general, our principles should cover attributions of the properties A and C to things other than ourselves: the probability that *this toss* will result in heads, given that *it* is a toss of the relevant type, should be such-and-such.

But now it matters how the relevant things are picked out. Suppose a certain coin will actually lands heads on its next toss. Then we can identify that toss as *the next toss of the coin resulting in heads*. The rational degree of belief in the hypothesis that the so-described toss will result in heads is 1, and remains 1 conditional on any hypothesis about chance. So Coordination fails. The problem here is that inadmissible information about the outcome has been smuggled into the way the chance process is picked out.

I will write '$R : A$' for the proposition that attributes the property A to the individual (event, process, etc.) identified by R. You may think of the *identifier* R as an individual concept or a Russellian definite description: '$R : A$' is true iff there is a unique individual that satisfies the condition R and this individual is A. I prefer to think of identifiers as binary relations, assuming with Lewis that the objects of credence are properties. When we attribute a property to individuals other than ourselves, we generally identify these individual by a relation they bear to us and our present location in space and time, e.g. when we consider the *next* toss of *this* coin. For any (binary) relation R and property A, $R{:}A$ is the property that applies to an object x iff there is a unique object R-related to x and this object has A. For the apparently non-relational way of picking out an individual as, say, 'the tallest man in the history of the universe', the identifier is the relation that holds between x and y iff y is the tallest man in the history of the universe of x. Even more degenerate cases are "singular" identifiers $=_\alpha$ which always pick out a particular individual α. $=_\alpha$ is the relation that holds between x and y iff y is α.

Now we can take the Principal Principle to say that for many ordinary identifiers R and propositions B,

$$P(R{:}A \;/\; R{:}C \wedge Ch_C(A)\!=\!x \wedge B) \approx x.$$

An identifier that picks out the next toss as 'next toss landing heads' would not count as suitable, because it contains inadmissible information.

The problem of inadmissible identifiers is rarely noted ([Skyrms 1980: 6ff.] is an exception), but it is a real phenomenon. For example, consider a variation of the Sleeping Beauty problem in which a second coin is tossed on Monday night. The proposition that *last night's coin* landed heads must then have different probability than the proposition that *the Sunday night coin* landed heads, even if last night's coin is the Sunday night coin (see [Schwarz 2012]). Inadmissible identifiers also often show up in the form of "observation selection effects". Suppose the chance of life to evolve on an Earth-like planet is 0.1. Conditional on this assumption, what is your subjective probability that life evolved on *the Earth-like planet on which you were born*? Not 0.1, of course. In this context, it is sometimes suggested that Coordination should be restricted to singular identifiers: if 'α' is a name of the Earth that conveys no qualitative information about its referent and its relation to ourselves, then conditional on the chance of life to evolve on an Earth-like planet being 0.1, the probability that life evolved on Earth-like planet α is plausibly 0.1 (see e.g. [White 2000]). I do not believe that this response works. For one thing, it is controversial whether it is possible to pick out individuals in a relevantly non-descriptive way. Moreover, even the singular information $=_\alpha{:}C$ can reveal inadmissible information, because it indicates that there are many Cs: in worlds where almost everything is C, it is less surprising that the individual α is C than in worlds where almost nothing is C; but the information that there are many Cs can be evidence about the outcome of a chance process under condition C.

To keep issues of admissibility out of the Coordination condition, I will resort to a technical trick. I will explain how to transform any credence function P into a function P_C relative to which a certain identifier ϵC picks out an individual of type C without conveying any further information about the individual or the world. Intuitively, $\epsilon C\!:\!A$ is the proposition that an arbitrarily chosen C is A, and P_C is an extension of P that believes in a corresponding process of arbitrary choice. More formally, let W be the class of elementary possibilities ("worlds") in the domain of the credence function P. Relative to each $w \in W$, any property F has an extension F_w, i.e. the class of F instances that exist relative to w. For any condition C, let W_C be the class of all pairs $\langle w, c\rangle$ where $w \in W$ and $c \in C_w$. It may help to picture W_C as generated from W much like centred worlds are generated from uncentred worlds by adding a 'you are here' arrow – except that the arrow now says 'the randomly selected C is here'. W_C adds an imaginary dimension to the space of possibilities that distinguishes between different Cs being "randomly selected". It also excludes all worlds where there are no Cs. Since ordinary, "real" propositions do not distinguish between which C is selected, a real proposition X shows up in W_C as the set of "complex" worlds $\langle w, c\rangle$ such that X is true at w. On the other hand, the complex proposition $\epsilon C\!:\!F$ is the set of $\langle w, c\rangle$ such that c is in the extension of F at w.

If W_C is finite, we can now define P_C as the probability distribution over W_C such that

$$P_C(\langle w, c\rangle) = \frac{P(w/\exists x Cx)}{|C|_w},$$

where $|C|_w$ is the number of Cs at w. Thus P_C conditionalises P on the assumption that there are Cs and then evenly divides the probability of any world $w \in W$ among all the pairs $\langle w, c\rangle$. This ensures that every C-instance in a world has equal probability of being "selected". It follows that for real propositions X,

$$(3) \qquad P_C(X) = P(X/\exists x Cx),$$

and that $P_C(\epsilon C\!:\!A)$ equals the P-expectation of the ratio of As among Cs:

$$(4) \qquad P_C(\epsilon C\!:\!A) = \mathbb{E}_P\left[\frac{|A \wedge C|}{|C|}\right]$$

These equalities can also be used to define P_C (to the extent that we need it) for cases where W_C is infinite, as long as zero probability is given to the hypothesis that there are infinitely many instances of C. I will return to this limitation at the end of the paper.

Now Coordination can be expressed as follows.

> **Coordination.**
> Normally, if P is a rational prior credence function, A and C are properties, and $P_C(Ch_C(A) = x) > 0$, then $P_C(\epsilon C : A / Ch_C(A) = x) \approx x$.

Informally: the extended subjective probability that an arbitrarily chosen C is A, given that the chance of A under C is x, should be approximately equal to x. Note that unlike in (1) and (2) above, there is no extra assumption about the instantiation of the condition C, since the randomly chosen individual of type C is already guaranteed to be an instance of C.

Coordination only indirectly links chance Ch_C to rational credence P by directly linking Ch_C to the C-transform P_C of P. Resiliency is also expressed in terms of P_C:

> **Resiliency.**
> Normally, if P is a rational prior credence function, A and C are properties, and $P_C(Ch_C(A) = x) > 0$, then there is a substantial range of propositions B and identifiers R such that $P_C(R : A / R : C \wedge Ch_C(A) = x \wedge B) = P_C(\epsilon C : A / Ch_C(A) = x)$.

$P_C(X)$ equals $P(X / \exists x Cx)$ if X does not involve the identifier ϵC. Thus the Principal Principle, combining Coordination and Resiliency, can be expressed directly in terms of P, as promised above:

> **Principal Principle.**
> Normally, if P is a rational prior credence function, A and C are properties, and $P(Ch_C(A) = x) > 0$, then there is a substantial range of identifiers R and propositions B such that $P(R : A / R : C \wedge Ch_C(A) = x \wedge B) \approx x$.

The present formulations reduce to those of section 1 if C is the tautologous property $\top$ and R is any non-defective identifier, since $P_C(R : A)$ is then equivalent to $P(A)$.

4 Russell's Argument

Now return to the Russellian argument from section 2. On the frequentist interpretation, $Ch_C(A) = x$ says that the relative frequency of As within a suitable sequence of Cs is x. Presumably this implies that the total number of Cs is positive and finite. Let P be a rational prior credence function for which $P(Ch_C(A) = x) > 0$. Let P' be P conditioned on $Ch_C(A) = x$. By (4), $P'_C(\epsilon C : A)$ is the expectation, by the lights of P', of the relative frequency of As among Cs. Since P' is certain that this ratio is x, $P'_C(\epsilon C : A) = x$. Hence $P_C(\epsilon C : A / Ch_C(A) = x) = x$. We've proved Coordination.

Thus far, all we needed was the assumption that rational prior credence obeys the probability calculus. However, Coordination is only half of the story. We also need

Resiliency. We have to show that there is (normally) a substantial range of ordinary propositions B and identifiers R such that $P_C(R : A/R : C \wedge B) \not\approx P(R : A)$, but $P_C(R{:}A/R{:}C \wedge Ch_C(A)=x \wedge B) = P_C(\epsilon C{:}A/Ch_C(A)=x)$. Together with Coordination, it then follows that $P(R{:}A/R{:}C \wedge Ch_C(A)=x \wedge B) = x$. Showing this requires more substantial assumptions.

Take a concrete example. Suppose the relevant C-instance is picked out demonstratively, say, as ‘the next toss of this coin’. If all you know is that the total ratio of heads among all tosses of the coin is 80%, what degree of belief should you assign to the hypothesis that the next toss will land heads? The probability calculus doesn’t settle the answer. You might be certain that the next toss lands tails, or heads, or give equal credence to heads and tails. But recall that we are talking about *prior* credence. If any of these attitudes are part of your priors, they are either based on no evidence at all, or on the information that the relative frequency is 80%. In this case, wouldn’t the attitude be irrational? If all you know is that the relative frequency of heads among some tosses is 80%, then you should be 80% certain that the next toss lands heads.

What’s at work here is the principle of indifference. Consider any random-looking sequence of heads and tails with 80% heads. If the length of the sequence is n, then there are n possibilities about the location of the “next toss”: it might be the first, or the second, . . . , or the nth. In the absence of relevant evidence, you should give equal credence to these n possibilities. It then follows that your credence in the next toss landing heads will be 0.8. The principle of indifference required here is closely related to the principle of induction. To be confident that the tosses you are going to observe land tails while the unobserved tosses mostly land heads would reflect an irrational, counterinductive attitude towards the world.

It is notoriously difficult to find a satisfactory, precise formulation of indifference, or of inductive probabilities more generally. But these difficulties should not cast doubt on the fact that the relevant constraints on prior credence exist. When a new galaxy is discovered, you should not be confident without evidence that it contains exactly 127 million planets; if ten people could have committed a murder, you should not be confident without evidence that it was the butler.

Fortunately, the present argument requires only a very restricted, compartmentalised application of indifference. It is not required that you distribute your credence uniformly over all ways things could be, which would presuppose a privileged parameterisation of logical space, and would arguably make it impossible to learn from experience (as can be seen by setting $\lambda = \infty$ in Carnap’s inductive logic). Imagine a grid imposed on logical space, each cell corresponding to a particular sequence of heads and tails. Set aside all cells in which the ratio of heads is not 80%, as well as possibilities in which the distribution of heads and tails shows a conspicuous pattern. Each of the remaining cells divides into subcells, corresponding to different possibilities about the location of the next toss. We require that within each of the cells from the original partition, you assign the same credence to every subcell. Nothing is said about how your credence should be divided between the larger cells, nor how it should be distributed within the subcells.

These remarks about *the next toss* carry over to other common identifiers such as *the previous toss* or *the toss presently reported by Jones*. They do not carry over to *the next toss that will land heads*. Here every specific hypothesis about a sequence of coin tosses still divides into different hypotheses about which of the tosses is the next toss that will land heads: the first, the second, and so on. But unless the sequence contains only heads outcomes, some of these subcells will be empty: if the ith element in the sequence is tails, then the ith element certainly isn't the next toss that lands heads. So your credence cannot be divided evenly between the subcells.

What about the extra information B? In the presence of the neutral identifier ϵC, it is easy to see that ordinary, real propositions are *always* admissible; i.e. for all real propositions B, $P_C(\epsilon C : A \wedge Ch_C(A) = x \wedge B) = x$. On the other hand, ordinary identifiers and ordinary propositions together can become inadmissible. If B suggests that Jones tends to report only tails outcomes, then your credence in the outcome reported by Jones being heads, conditional on the relative frequency being 80% and B, won't be 0.8. We could say that in the presence of B, the identifier *the toss reported by Jones* is inadmissible. Or we could say that in the presence of this identifier, the information B is inadmissible. It doesn't really matter. (Technically, the addition of B in the Resiliency condition and the Principal Principle is redundant, since the information in B can always be folded into the identifier R.)

In section 2, we saw that information B about previous outcomes, combined with information about the total length of a sequence, can be inadmissible for frequentist chance. On the other hand, we also saw that if the credence function is not unduly opinionated, then Resiliency holds with respect to all B that specify not too many other outcomes. The relevant constraint on rational credence – exchangeability among random sequences – is another highly restricted form of indifference. As Lewis [1994: 229] points out, frequency information also tends to screen off many other facts that would otherwise be relevant to the outcome of the next toss, such as symmetries or asymmetries in the coin and the tossing procedure.

5 Best System Probabilities

Let us move on to the leading Humean theory of chance: the best system approach. Here, 'chance' is defined indirectly via statistical theories. Let a *theory* be any logically closed set of sentences in a suitable language that includes resources to express probabilities. Given the total history H of (relevant) events in a world, theories can be ranked by their simplicity, strength, fit and possibly further criteria. Then chance is defined as the probability function employed in whatever theory ranks highest, on balance, in terms of these virtues. (See [Lewis 1994], [Loewer 2004], [Hoefer 2007] for more detailed expositions, and different ways of filling in the details.)

The *fit* between a theory T and a history H measures the extent to which T assigns high probability to events in H. Lewis suggested that if P^T is the probability function specified by theory T, then $P^T(H)$ can serve as measure of fit. But this presupposes that statistical theories assign an absolute, unrelativised probability to complete histories. A natural generalisation to the present framework would use the product of $P^T_C(A)$ for each occurrence of an outcome A under a condition C in the history:

$$\text{(5)} \qquad \prod_{\langle C,A\rangle \in H} P^T_C(A).$$

Here I imagine that a history is a sequence (or a multiset) of condition-outcome pairs.

Formally, (5) defines a probability distribution over any space of histories H which agree in their frequency distribution over conditions $C \in \Gamma$. To make this more explicit, let T be any theory. Partition the space of histories by the distribution of frequencies over T's conditions Γ: $H \sim H'$ iff $|C|_H = |C|_{H'}$ for all $C \in \Gamma$. For any cell F in this partition, define $\mathit{Fit}_{T,F}$ as the probability distribution over F given by

$$\mathit{Fit}_{T,F}(H) = \prod_{\langle C,A\rangle \in H} P^T_C(A).$$

$\mathit{Fit}_{T,F}$ resembles the "extended" frequentist chance function of section 2, but here it is not meant to represent T's probability for a history.

What actually matters for the fit of a history to a theory are only the frequencies of outcomes in the history, not their order:

$$\mathit{Fit}_{T,F}(H) = \prod_{C\in\Gamma} \prod_{A\in\Omega_C} P^T_C(A)^{|C\wedge A|_H}.$$

Thus we can also measure fit directly in terms of a history's frequency distribution. Let $\langle k_1, \ldots, k_n\rangle_C$ be the set of histories $H \in F$ for which the outcomes $A_1, \ldots, A_n$ under condition C have frequency $k_1, \ldots, k_n$, respectively. $\mathit{Fit}_{T,F}(\langle k_1, \ldots, k_n\rangle_C)$ is given by the multinomial formula

$$\mathit{Fit}_{T,H}(\langle k_1, \ldots, k_n\rangle_C) = \binom{|C|_H}{k_1, \ldots, k_n} \prod_{i=1}^{n} P^T_C(A_i)^{k_i}.$$

The fit of a history (in F) is the product of these values, for every condition C.

An alternative way to measure fit is to look at the differences $\Delta = |C \wedge A|_H - P^T_C(A)|C|_H$ between the "observed" frequencies $|C \wedge A|_H$ of A outcomes under condition C in a history and the "expected" frequencies by the light of the theory, $P^T_C(A)|C|_H$. Intuitively, the more the observed frequencies match the expected frequencies, the better the fit between theory and history. Aggregating the (squared normalised) differences Δ for all outcomes A under all conditions C yields

$$X^2 = \sum_{C \in \Gamma} \sum_{A \in \Omega_C} \frac{(|C \wedge A|_H - P^T_C(A)|C|_H)^2}{P^T_C(A)|C|_H}.$$

The lower X^2, the better the fit.

On reflection, this measure is only plausible if the relevant frequencies are reasonably large. In this case, the X^2 value of a history H can be converted into an approximation of $\mathit{Fit}_{T,F}(H)$, so the two measures of fit are not really alternatives: since $\mathit{Fit}_{T,F}(\Delta = x)$ then follows an approximately normal distribution, $\mathit{Fit}_{T,F}(X^2 = x)$ approaches a sum of squared standard normal distributions; the χ^2 function with $\sum_{C \in \Gamma}(|\Omega_C| - 1)$ degrees of freedom thus yields an approximation of the $\mathit{Fit}_{T,F}$ of the set of histories in which the frequencies are at least as far from the expectation as in H. (The reasoning here parallels the reasoning behind the χ^2 test for "goodness of fit" in frequentist statistics.)

Now let P' be a rational prior credence function conditioned on the assumption that (i) the frequency of Cs in actual history is k, and (ii) the best theory T assigns probability x to outcome A under condition C. Note that whatever the total frequency distribution F and best theory T might be, the distribution of $\mathit{Fit}_{T,F}$ over values of $|C \wedge A|$ is a binomial with mean xk. Arguably, the P'-expectation of $|C \wedge A|$ should equal this mean xk:

(6) $\mathbb{E}_P(|C \wedge A| \;/\; |C| = k \wedge Ch_C(A) = x) = xk.$

It follows that the P'-expectation of the relative frequency $\frac{|C \wedge A|}{|C|}$ equals x. By definition, this expectation equals $P'(\epsilon C\!:\!A)$; so

(7) $P(\epsilon C\!:\!A \;/\; |C| = k \wedge Ch_C(A) = x) = x.$

Since (7) is true for all k, we get the strict Coordination rule

(8) $P(\epsilon C\!:\!A \;/\; Ch_C(A) = x) = x.$

The only substantial assumption here is (6). Where does that come from? Recall that the information $Ch_C(A) = x$ says that the best theory assigns probability x to A under C. Part of what makes a theory good is its fit, and the closer the actual frequency $|C \wedge A|$ to the expectation $P^T_C(A)|C|$, the better the fit. Since there is also a simplicity constraint on theories, the actual frequency may come apart from its expectation: it may be too high, or too low. (6) assumes that your rational prior credence in either sort of deviation balances out so that the expected deviation is zero.

Picture the binomial curve for $\mathit{Fit}_{T,F}$ over possible frequencies of A. If all you know is that the history has reasonably high fit, then where do you think the A frequency lies under the goodness of fit curve? Arguably, you should believe that it is not too far from the maximum of the curve, and your credence in deviations on either side should balance out.

Again, the rationality constraint reflected in (6) is a restricted principle of indifference. A sufficient (but not necessary) constraint to get (6) would be to assign uniform prior credence to different hypotheses about the relative frequency of As under condition C. This is precisely the assumption Laplace used for his famous derivation of the rule of succession. By contrast, assigning uniform credence to all histories would not support (6). For example, if a die is tossed six times, then there are more histories with two sixes than histories with one six or zero sixes (15 vs. 6 vs. 1). Relative to a theory that assigns uniform probability 1/6 to each outcome of a die toss, the single history with zero sixes has somewhat better fit than the histories with two sixes (0.33 vs. 0.2). If you start out with a uniform prior over histories, you will be more confident, conditional on the best theory treating the die as fair, that the actual history has two or more sixes than that it has none, so (6) fails. But indifference between all histories is irrational: it would make it impossible to learn from experience.

On the other hand, there are cases in which (6) may actually be false. Suppose the best system assigns probabilities to informationally very rich outcomes, e.g. to the hypothesis A that the universe contains precisely 1000 coin tosses all of which result in heads. The information that the best system assigns low probability x to A may then imply that A is false, since the best system of a world where A is true would not treat the coin tosses as chancy at all. So the P'-expectation of the frequency of A is zero, rather than xk. This is the phenomenon of undermining. As [Lewis 1980: 111f.] points out, Humean accounts of chance that allow for undermining propositions like A are incompatible with the strict Coordination condition (i.e., Coordination without the softeners). It is therefore unsurprising that the present derivation of (8) breaks down for such propositions.

What about Resiliency? Here, most of what I said for frequentism carries over. In fact, the best system account generally yields a wider domain of resiliency. That's because best system probabilities have to fit many frequencies, for many conditions. For example, if the state transitions for a certain system are modeled as a random walk, then the relative frequency of transitions from state C to state A may differ widely from the best system's probability $P_C(A)$ – especially if C or A is rare. Hence information about previous transition frequencies has little effect on how likely you should deem a transition from C to A, once you know the chance.

I have assumed that histories and outcome spaces are finite. If we lift this assumption, we run into the "zero-fit" problem (see [Elga 2004]). There are really two problems here, one arising from infinite outcome spaces, and one from infinite histories. Infinite outcome spaces are common in science, because outcomes are often real-valued. There are several ways to accommodate this in goodness of fit measures. A common method in statistics is to replace individual outcomes by reasonably chosen intervals, for example by partitioning the possible outcomes into $\sqrt{|C|}$ many intervals with uniform expected frequency.

Infinite histories are harder to deal with. If there are infinitely many instances of condition C in a history, our goodness of fit measure will no longer distinguish better from worse theories, since they all have zero fit. We also run into problems with the definition of P_C. A simple way around these issues might be to focus on finite subsets of Cs. If the world contains infinitely many Cs, we can look at increasingly large "samples", including all Cs within a certain distance from ourselves. If the world is well-behaved, the relative frequencies in these samples, and thereby the order of theories by fit, should converge. Of course, there is no logical guarantee that the world is well-behaved, but ill-behaved worlds deserve little rational credence, especially conditional on the hypothesis that the best system specifies probabilities relative to C.

References

Robert Black [1998]: "Chance, Credence and the Principal Principle". *British Journal for the Philosophy of Science*, 49: 371–385

Jonathan Cohen and Craig Callender [2009]: "A Better Best System Account of Lawhood". *Philosophical Studies*, 145(1): 1–34

Bruno de Finetti [1937]: "La Prévision: ses lois logiques, ses sources subjectives". *Annales de l'Institute Henri Poincaré*, 7: 1–68

David Deutsch [1999]: "Quantum Theory of Probability and Decisions". *Proceedings of the Royal Society of London*, A455: 3129–3137

Persi Diaconis and David Freedman [1980]: "De Finetti's Generalizations of Exchangeability". In R. Jeffrey (Ed.) *Studies in Inductive Logic and Probability,* vol 2, chapter 11. Berkeley: University of California Press

Adam Elga [2004]: "Infinitesimal Chances and the Laws of Nature". In Frank Jackson and Graham Priest (Eds.) *Lewisian Themes: The Philosophy of David K. Lewis,* Oxford: Oxford University Press, 68–77

Roman Frigg and Carl Hoefer [2010]: "Determinism and Chance From a Humean Perspective". In Friedrich Stadler et al. (Ed.) *The Present Situation in the Philosophy of Science,* Springer, 351–72

Ned Hall [1994]: "Correcting the Guide to Objective Chance". *Mind*, 103: 505–517

— [2004]: “Two Mistakes about Credence and Chance”. *Australasian Journal of Philosophy*, 82: 93–111

Carl Hoefer [2007]: “The Third Way on Objective Probability: A Skeptic’s Guide to Objective Chance”. *Mind*, 116: 549–596

Colin Howson and Peter Urbach [1993]: *Scientific Reasoning*. La Salle (Ill.): Open Court Press, 2nd edition

David Lewis [1979]: “Attitudes *De Dicto* and *De Se*”. *The Philosophical Review*, 88: 513–543

— [1980]: “A Subjectivist’s Guide to Objective Chance”

— [1994]: “Humean Supervenience Debugged”. *Mind*, 103: 473–490

Barry Loewer [2004]: “David Lewis’s Humean Theory of Objective Chance”. *Philosophy of Science*, 71: 1115–1125

Christopher J. G. Meacham [2005]: “Three Proposals Regarding a Theory of Chance”. *Philosophical Perspectives*, 19(1): 281–307

David H. Mellor [1971]: *The Matter of Chance*. Cambridge: Cambridge University Press

Bertrand Russell [1948]: *Human Knowledge: Its Scope and Limits*. London: George Allen and Unwin

Wolfgang Schwarz [2012]: “Lost memories and useless coins: revisiting the absendminded driver”. Manuscript

Brian Skyrms [1980]: *Causal Necessity. A Pragmatic Investigation of the Necessity of Laws*. New Haven: Yale University Press

Michael Strevens [1999]: “Objective Probability as a Guide to the World”. *Philosophical Studies*, 95: 243–275

Richard von Mises [1919]: “Grundlagen der Wahrscheinlichkeitsrechnung”. *Mathematische Zeitschrift*, 5: 52–99

David Wallace [2012]: *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford: Oxford University Press

Roger White [2000]: “Fine-Tuning and Multiple Universes”. *Noûs*, 34(2): 260–8211