You Can't Always Get What You Want*
Some considerations regarding conditional probabilities

Wayne C. Myrvold
Department of Philosophy
The University of Western Ontario
wmyrvold@uwo.ca

August 22, 2013

Abstract

The standard treatment of conditional probability leaves conditional probability undefined when the conditioning proposition has zero probability. Nonetheless, some find the option of extending the scope of conditional probability to include zero-probability conditions attractive or even compelling. This articles reviews some of the pitfalls associated with this move, and concludes that, for the most part, probabilities conditional on zero-probability propositions are more trouble than they are worth.

*But if you try, sometimes, you find you get what you need.

1 Introduction

Let $\mathcal{A}$ be a set of propositions, closed under Boolean operations, let Pr be a probability function on $\mathcal{A}$, and, for some proposition C, let Pr_C be another probability function on $\mathcal{A}$, to be thought of as yielding probabilities conditional on C.[1] It is uncontroversial that, if C is in $\mathcal{A}$, these should satisfy

$$Pr(AC) = Pr_C(A)\, Pr(C). \tag{1}$$

If $Pr(C) > 0$, then the unconditional probability function Pr, together with the requirement that (1) hold, uniquely determines $Pr_C(A)$, for any $A \in \mathcal{A}$:

$$Pr_C(A) = \frac{Pr(AC)}{Pr(C)}. \tag{2}$$

If, however, $Pr(C) = 0$, then $Pr(AC)$ is also equal to zero, and (1) is satisfied for any value whatsoever of $Pr_C(A)$, and so (1) leaves $Pr_C(A)$ completely undetermined.

One reaction, the standard one, is to leave $Pr_C(A)$ undefined except for $C \in \mathcal{A}$ with $Pr(C) > 0$. But, since (1) places no constraints whatsoever on the function Pr_C when $Pr(C) = 0$, for such propositions we are free, without fear of violating this condition, to define Pr_C to be any probability function whatsoever on $\mathcal{A}$. Instead of relying on (1) to define conditional probability functions in terms of the unconditional probability function Pr, we can take conditional probability as primitive. This is a route that has been recommended by a number of authors over the years. In support of this, cases are sometimes adduced that suggest that there are probabilities conditional on zero-probability propositions that have clearly defined values (see §2, below). Moreover, it seems we *have* to regard some probabilities conditional on zero-probability propositions as well-defined, in order to do justice to statistical practice, since statistical practice invokes likelihood functions, which ascribe probabilities to data as a function of some continuously varying parameter, and these are well-defined for all parameter values even if every point value of the parameter is ascribed zero probability. We do not want to eschew the use of such functions; does this not commit us to probabilities conditional on zero-probability propositions?

In this essay, I hope to convince the reader that things are not so straightforward. The examples that purport to show that there are clear-cut answers to requests for probabilities conditional on propositions of probability zero are misleading. We can *give* such questions answers by requiring that the conditional probability functions possess certain symmetry properties, but this is our choice, not dictated by the nature of the problem, and we should not let the intuitive appeal of such symmetry properties blind us to the fact that we must stipulate that the conditional probabilities have them, in order for the questions to acquire determinate answers. Moreover, there will be cases in which symmetry conditions that we may wish to impose will clash with each

[1] We will also use the notation $Pr(A|C)$, when convenient.

other, or may clash with the *desideratum* of countable additivity, illustrating Jagger's Theorem: You Can't Always Get What You Want.[2]

Furthermore, the consequences of taking the standard route, and leaving undefined probabilities conditional on null propositions (that is, propositions of unconditional probability zero), are not as dire as some would make them out to be. Though there are cases (such as the likelihood functions already mentioned) in which quantities appear that can unproblematically be taken to be probabilities conditional on null propositions, they need not be, and the theory goes along straightforwardly if we take all conditional probabilities to have conditions with positive probability.

Some will be undaunted, and will insist on introducing a host of null-condition conditional probabilities. This can be done, but, if it is done, it should be *done* and not merely gestured at: those who invoke probabilities conditional on null propositions should specify which pairs of propositions A, B they take the conditional probability $\Pr(A|B)$ to be defined for, and specify the values of these conditional probabilities.

2 Examples

Consider the following examples.

Example 1.[3] A number is chosen, with uniform probability, from the interval $[0, 1]$. Conditional on the supposition that the chosen number is either $1/4$ or $3/4$, what is the probability that it is $1/4$?

Example 2. (Borel-Kolmogorov).[4] A point is chosen, with uniform probability, on the surface of the earth, which we treat as a perfect sphere.

a). What is the probability that the chosen point is in the Western Hemisphere, given that that it lies on the equator?

b). Conditional on the chosen point lying on the great circle containing the Greenwich meridian, what is the probability that it lies closer to the equator than to a pole?

[2]The fact that, in probability theory, we can't always get what we want, is a familiar fact. It is well-known that there are no probability functions satisfying certain symmetry conditions, countable additivity, and the *desideratum* of having the probability function defined on arbitrary subsets of our sample space. Consider, for example, the task of defining a uniform distribution—that is, a distribution invariant under all rotations—on the unit circle. There can be no distribution that is invariant under rotations, is countably additive, and is defined on all subsets of the unit circle. The proof is found in many probability texts, *e.g.* Billingsley (2012, p. 47). The standard response is to preserve countable additivity and to restrict the domain of definition of the probability function to the measurable sets. If, however, one is willing to give up countable additivity, it *is* possible to extend the probability function to one defined on arbitrary subsets; as Banach (1923) showed, in one and two dimensions it is possible to extend the Lebesgue measure to a finitely additive measure defined on all subsets that is invariant under transformations that preserve distances. The well-known Banach-Tarski paradox shows that this is impossible in three-dimensional space.

[3]Adapted from Hájek (2003).

[4]Based on Kolmogorov (1950, §V.2). See also Jaynes (2003, §15.7), Hájek (2003, §4.4).

Example 3. A number is chosen, with uniform probability, from the interval $[0, 1]$. Conditional on the supposition that the chosen number is rational, what is the probability that it is greater than $1/2$?

For many, perhaps most, readers, each of the above questions will have an obvious answer. This should give us pause. In each case, an unconditional probability distribution is given, and a question asked about probabilities conditional on a proposition with probability zero. As already emphasized, the unconditional probability distribution, though it does not prohibit such questions from having answers, as it will not clash with any answers that we give, by the same token cannot determine what those answers must be. Any intuition that these questions have determinate answers must involve some tacit assumption not contained in the set-up of the problem. We should endeavour to make these tacit assumptions explicit, and explore the consequences of requiring our conditional probabilities to satisfy them.

3 Symmetry Conditions

3.1 Example 1.

Example 1 seems beguilingly simple. It may seem that the symmetry of the problem dictates the answer $1/2$, on pain of irrationality. Nothing at all in the set-up of the problem favours either $1/4$ or $3/4$.

But consider this variant on the question. Suppose that the number is chosen from the unit interval, not with uniform distribution, but according to a distribution given by the density function

$$f(x) = 2x. \tag{3}$$

Now ask the question: conditional on the number chosen being either $1/4$ or $3/4$, what is the probability that it is $1/4$?

Here, I suspect, intuitions will vary. To some, the answer might still be, obviously, $1/2$. To others, reflecting on the fact that the number chosen is more likely to be greater than $1/2$ than less than $1/2$, will regard $3/4$ as the more probable value. This intuition can be given a numerical value by considering, that, for any sufficiently small positive ϵ,

$$\frac{Pr(X \in [\frac{3}{4} - \epsilon,\ \frac{3}{4} + \epsilon])}{Pr(X \in [\frac{1}{4} - \epsilon,\ \frac{1}{4} + \epsilon])} = 3, \tag{4}$$

which suggests that the number $3/4$ is 3 times as probable as $1/4$.

Suppose, now, that we change the question only slightly, and ask, if the number is chosen from the unit interval according to a distribution with density (3), what is the probability, conditional on the number chosen being either $1/2$ or $\sqrt{3}/2$, that it is $1/2$? Similar considerations suggest that $\sqrt{3}/2$ is more probable than $1/2$.

If we give this answer, we have thereby achieved incoherence, because this last question is just our first question rephrased. X being uniformly distributed on the unit

interval is the same as $\sqrt{X}$ being distributed with density (3), and so, choosing X with uniform distribution, and asking whether 3/4 is more probable than 1/4, conditional on the chosen number being one of the two, is the same as choosing $\sqrt{X}$ according to (3) and asking whether $\sqrt{3}/2$ is more probable than 1/2.

We can escape incoherence by adopting the convention that, if a number is chosen according to any probability distribution on the unit interval, then, for any finite subset of the unit interval, the probability conditional on the number being in that subset is the same for every member of the set. And, if we are to have equiprobability when the distribution is uniform, this is the *only* way to escape incoherence, since, for any probability distribution that is yielded by a density function, there will always be *some* random variable that is uniformly distributed. But this convention may seem odd to some: consider a density function that is very sharply peaked around 1/2. On the convention under consideration, conditional on the supposition that the chosen number is either 1/2 or 9/10 (which could be as many standard deviations away from the peak as we like), 1/2 and 9/10 are equally probable.

These considerations will, I hope, lead any readers who initially regarded question 1 as having an obvious answer to conclude: things aren't as straightforward as they seemed.

3.2 The Sphere

Consider, again, example 1. A point is chosen, with uniform probability, on the surface of a sphere, and we are asked to reflect on the questions: a) What is the probability that the chosen point is in the Western Hemisphere, given that that it lies on the equator? b) Conditional on the chosen point lying on the great circle containing the Greenwich meridian, what is the probability that it lies closer to the equator than to a pole?

For question 1(a), the seemingly obvious answer is 1/2. For 1(b), the obvious answer might seem to be 1/2, again, as half of the length of any meridian consists of points that are closer to the equator than to a pole.

But consider this: it is *not* true that 1/2 of the earth's surface is closer to the equator than it is to a pole; more of it is closer to the equator. The probability that a point chosen with uniform probability is closer to the equator than to a pole is $1/\sqrt{2} \approx 0.707$. Since every point lies on some meridian, we might want to say that the probability, conditional on our point lying on the Greenwich (or any other) meridian, of being closer to the equator than to a pole, is $1/\sqrt{2}$.

Any reader who is wondering whether the *correct* answer to 1(b) is 1/2 or $1/\sqrt{2}$ or some other number is reminded: the setup of the problem does not determine *any* answer. The answer of 1/2 seems to rely on some intuition that the conditional probabilities should share relevant symmetries with the unconditional distribution. An intuition is a dangerous thing; we would do well to replace the intuition with an explicit requirement regarding symmetries.

If a probability space $\langle S, \mathcal{A}, Pr \rangle$ is invariant under a transformation T, then *ipso*

facto so is the standard conditional probability space[5] $\langle S, \mathcal{A}, \mathcal{A}^*, P^* \rangle$. We may want to use symmetry considerations to extend the standard conditional probability space to one that includes conditionalization on null propositions. As a first pass, we might be tempted to require that our conditional probability space be invariant under all transformations that leave the unconditional probability space that we started with invariant; this, we might speculate, is the requirement needed to underwrite the "obvious" answer to question 1(a). A moment's reflection, however, reveals that this is unreasonably strong. Let C be any probability-zero subset of S, and, for *any* transformation T_C of C, consider a transformation of $\langle S, \mathcal{A} \rangle$ that consists of performing T_C on C and doing nothing elsewhere. Since $Pr(C) = 0$, this transformation does not change the unconditional probability of any set. Thus, to require invariance under arbitrary transformations that leave the unconditional probability space invariant entails that probabilities conditional on a null set C be invariant under arbitrary transformations of C, a requirement that is satisfiable when C is a finite set but not otherwise.

The set S of events might have additional structure that we can require our transformations to preserve. In the sphere case, the elementary events are choices of points on a sphere, and these points have distances between them. We can restrict our attention to transformations of our probability space that preserve these distances. These are just the rigid rotations of the sphere. Requiring invariance under all rigid rotations entails that the conditional probability function, conditional on the chosen point lying on a circle, be invariant under the subgroup of rotations that leave the circle invariant. This is is uniquely satisfied by a uniform distribution on the circle.

If the intuition that the obvious answer to the sphere questions 1(a) and 1(b) is 1/2 rests on an implicit assumption that probabilities, conditional on the point lying on a circle, should be invariant under rotations that leave the circle invariant, then, rather than leave this implicit, we should place it as an explicit condition on our conditional probability space. Can we do this? If we're not too demanding about the extent of the set $\mathcal{B}$ on which we conditionalize, then it is easy to show that we can. This is done in Appendix 3, where we construct a conditional probability space that includes conditionalization on all circles and subsets of circles of nonzero length, and is invariant under rigid rotations of the sphere.

We might want more than this in our domain of conditionalization. Can our conditional probability space be extended in such a way that it includes conditionalization on *all* measurable subsets of the sphere, and preserves symmetry under rotations?

If we demand countable additivity, then the answer is easy: no, we can't. Given a coordinatization of the sphere by latitude and longitude, consider E_Q, the set of points on the equator whose longitudes are rational numbers. This set is invariant under rational rotations of the sphere about its axis. Invariance under such rotations requires that the probability, conditional on E_Q, ascribed to any interval of the equator be proportional to the length of the interval, and this in turn requires the probability assigned to single points on the equator to be zero. But $P_{E_Q}(E_Q)$ must be equal to one, and so the conditional probability function P_{E_Q} cannot be countably additive.

[5] See Appendix 1 for definitions.

Similar considerations apply, of course, to our Example 3. Our unconditional probability function is invariant under translations of the unit interval (modulo 1). The set of rationals in the unit interval is invariant under the subgroup consisting of translations through a rational distance. Imposing translation symmetry on probabilities conditional on the number chosen being rational gives the expected answer: conditional on the number being rational, the probability that it lies in any interval is equal to the length of that interval. But this comes at the cost of violating countable additivity. If we conditionalize on the rationals we are faced with a choice between a symmetry condition that may be desired, and preserving countable additivity. This is something that we do not have to face when conditioning on sets of nonzero probability; if Pr is countably additive, and $Pr(C) > 0$, then Pr_C is also countably additive.

Suppose we're willing to give up countable additivity. Is there a conditional probability space that permits conditionalization on arbitrary measurable subsets of the sphere, and is invariant under rotations? Since this will include conditionalization on measure-zero subsets of S that are neither invariant under rotations nor contained in nontrivial subsets that are invariant under rotations, it is likely that, if such conditional probability spaces do exist, rotational symmetry will not suffice for uniqueness. We should expect that, if there are any, there are many such spaces, and that it would not be a trivial task to specify one. It is, as far as I know, an open question whether such conditional probability spaces exist. Philosophers who write as if one can blithely assume that such conditional probability spaces exist are kindly requested to show that they do, and, if there is more than one, to specify which one they have in mind.

3.3 The Eternal Coin

In the case of the sphere, things worked out (reasonably) well. We were able to identify a natural group of symmetries, and imposition of these symmetries entailed one of the 'obvious' answers to our questions. In other cases, we will not be so lucky. Symmetries that we may wish to impose can come into conflict.

An interesting example of this is provided by Cian Dorr (2010), in the set-up that he calls "The Eternal Coin." The Eternal Coin is a fair coin that is flipped every day, throughout an infinite past, and will continue to be flipped every day into an infinite future. In the absence of any other information about the coin, we are invited to consider credences in propositions such as

H: The Coin lands Heads today.
P: The Coin landed Heads on every day in the past.
F: The Coin will land Heads on every day in the future.

All probabilities—including probabilities conditional on propositions with probability zero—will be taken to be predicated on the setup being as we have described it.[6]

[6]This is necessary because, if one has nonzero credence that the coin is not fair, or that the tosses are not independent, then conditionalization on either F or P will send credence that the setup is as described

We construct a probability space as follows. Our set Υ of elementary events is the set of bi-infinite sequences of Heads and Tails. To form a σ-algebra $\mathcal{F}$ of measurable sets, we proceed as follows. For any finite set of integers K, and any $u \in \Upsilon$, we form a *cylinder set* $C_K(u)$ consisting of all elements of Υ that agree with u on the set K. That is, a cylinder set is the set of all events that agree on some finite subset of integers. We take $\mathcal{C}$ to be the smallest σ-algebra containing all cylinder sets.

To define a probability measure Pr on $\langle \Upsilon, \mathcal{C} \rangle$, it suffices to specify the probabilities of cylinder sets. To do this, we assign, for any k-element set K, the probability 2^{-k} to each cylinder set $C_K(u)$. This function has a unique countably additive extension to $\mathcal{C}$, which we will take to be our probability measure Pr. This gives us a probability space $\langle \Upsilon, \mathcal{C}, Pr \rangle$.

This probability measure has, as expected, the following features:

a). Each individual flip has equal probability 1/2 for H and T.

b). Outcomes of distinct flips are independent: if K, L are disjoint sets, then, for all $u, v \in \Upsilon$,

$$Pr(C_K(u) \cap C_L(v)) = Pr(C_K(u)) \cdot Pr(C_L(v)).$$

For any set of integers L, let $F_L : \Upsilon \to \Upsilon$ be the 'bit flip' transformation on L, that is, the transformation that consists of exchanging H and T at each place in L. Our probability space is invariant under all such transformations.

Our probability space is also invariant under permutations of the integers. For any bijection $\pi : \mathbb{Z} \to \mathbb{Z}$, let $\mathsf{T}_\pi : \Upsilon \to \Upsilon$ be the operation whose action on a bi-sequence u permutes the values of u,

$$(\mathsf{T}_\pi u)_k = u_{\pi(k)}. \tag{5}$$

Permutations that will be of particular interest are the shift operations. For any integer n, let $\mathsf{S}_n : \Upsilon \to \Upsilon$ be the operation of shifting everything n places:

$$(\mathsf{S}_n u)_k = u_{k-n}. \tag{6}$$

Invariance under shift operations means that, although our coordinatization has a distinguished origin (the day 0, which we are calling "today"), our probability space is invariant under shift of this origin.

If P_n is the proposition that the coin landed Heads on the past n days, then $Pr(P_n) = 2^{-n}$. Since P entails P_n for each n, it follows that $Pr(P) = 0$. Similarly, $Pr(F) = 0$.

The function Pr, of course, uniquely determines probabilities conditional on propositions with non-zero probability. Dorr invites us to consider probabilities conditional on some zero-probability propositions, such as P, F, and $P \vee F$. It is, of course, possible to extend our probability assignments to include probabilities conditional on propositions such as these, and this can be done in a variety of ways.

to zero.

Here's one way to do it. For any n, let $K_n = [-n, n]$, and, for any $A \in \mathcal{C}$, let A_n be the proposition that commits only to what A says about coin flips in K_n, and says nothing about what happens outside this interval. That is, take

$$A_n = \bigcup_{u \in A} C_{K_n}(u). \tag{7}$$

Our set $\mathcal{B}$ of conditions will consist of all $B \in \mathcal{C}$ for which there exists N such $Pr(B_n) > 0$ for all $n > N$. For $B \in \mathcal{B}$, let $\mathcal{A}_B$ be the set of $A \in \mathcal{C}$ such that the sequence $Pr(A_n|B_n)$ converges to a limit as $n \to \infty$, and, for $A \in \mathcal{A}_B$, take

$$Pr(A|B) = \lim_{n \to \infty} Pr(A_n|B_n). \tag{8}$$

A few of the conditional probabilities that we thereby obtain are,

$$\begin{aligned}
&Pr(H|F) = Pr(T|F) = Pr(H|P) = Pr(T|P) = 1/2;\\
&Pr(P|P \vee F) = Pr(F|P \vee F) = 1/2;\\
&Pr(P|P \vee HF) = Pr(F|HP \vee F) = 2/3;\\
&Pr(HF|P \vee HF) = Pr(HP|HP \vee F) = 1/3;\\
&Pr(P \vee HF|P \vee F) = Pr(HP \vee F|P \vee F) = 3/4.
\end{aligned} \tag{9}$$

The limiting procedure we have sketched is, of course, only one possible limiting procedure, and no claim is made for priority of this over other procedures. We have made a frankly arbitrary choice, and have obtained the above conditional probabilities; other choices will yield other values.

The conditional probabilities we have obtained preserve independence and bit-flip symmetry. The limiting procedure we have chosen manifestly breaks shift symmetry. Unsurprisingly, the conditional probabilities we obtain from it also violate shift symmetry. To see this, consider the one-day shift S_1. We have,

$$\mathsf{S}_1(P) = HP \qquad \mathsf{S}_1(HF) = F \tag{10}$$

However,

$$Pr(P|P \vee HF) \neq Pr(HP|HP \vee F) \tag{11}$$

We therefore have extended our probability function in a way that respects independence of distinct flips, and also bit-flip symmetry, but violates shift symmetry. We should ask whether we can do better, and extend our probability function in such a way that all of the above conditional probabilities are defined so as to respect all of these symmetries.

Dorr shows that, counterintuitively,[7] the answer is no. Provided that $Pr(P|P \vee F)$ and $Pr(F|P \vee F)$ are defined and are both positive, shift invariance entails that

$$Pr(H|F) = Pr(H|P) = 1. \tag{12}$$

Proof is given in Appendix 3.

A similar argument yields a violation of countable additivity. Let P^+ be the proposition that the coin has landed Heads every day in the past but will land Tails sometime, either today or in the future, and let F^+ be the proposition that the coin will land Heads every day in the future, but landed Tails today or sometime in the past. Shift invariance, together with the conditions that $Pr(P^+|P^+ \vee F^+)$ and $Pr(F^+|P^+ \vee F^+)$ are defined and are both nonzero, entails that, for every n, the probability conditional on P^+ that the coin lands Heads today and every day for n days into the future is one. This in turn entails (letting H_n be the proposition that the coin will land Heads n days from now and T_n, the proposition that it will land Tails), that, for each n.

$$\begin{array}{l} Pr(H_n|P^+) = 1; \\ Pr(T_n|P^+) = 0; \end{array} \tag{13}$$

even though the probability, conditional on P^+, that, for some n, T_n is true, is unity.

Including the propositions P, F, and $P \vee F$ in the set of propositions on which we can conditionalize, and imposing shift symmetry, is possible, but it comes at a high cost: we lose independence; it is no longer true that conditionalization on a proposition that specifies outcomes on a set of days not including today leaves the probability of the coin landing Heads today unchanged. Symmetry conditions that we would like our conditional probability space to respect clash; we can't get all that we want.

Depending on our purpose, we might prefer to preserve one or the other of the symmetries. If the Eternal Coin is being considered as an idealization of a situation in which a coin is tossed a large but finite number of times, then shifts will not be symmetries of the finite system, which is our real object of interest, and so it will not be important for our purposes to demand shift invariance of the conditional probability space. There might be other purposes for which shift invariance is of such paramount importance that it would be worth abandoning independence (though it is hard to see why it would not be preferable to simply leave those conditional probabilities undefined).

If we think of the setup as involving an actual bi-infinite sequence of coin tosses, not an idealization of a finite set-up, then, as Dorr convincingly argues, violation of shift invariance is bizarre. Dorr invites us to imagine ourselves causally isolated from the Eternal Coin. I learn nothing about the outcomes of its flips as the days pass. Now, consider the following: HP, the proposition that the coin lands Heads today and landed Heads every day in the past, is the proposition that, tomorrow, I will express by the words, "The coin landed Heads every day in the past," the same sentence that

[7]Perhaps. The more one thinks about what is required to give values to these conditional probabilities, the less clear it becomes that we have intuitions about them at all.

I use today to express the proposition P. Similarly, $HP \vee F$ is the proposition that I will express tomorrow using the same words I use today to express $P \vee HF$. Today, when I say "My credence that the coin landed every day in the past, conditional on the supposition that it either landed Heads every day in the past or will land Heads today and every day in the future," I denote $Pr(P|P \vee HF)$; tomorrow, the same phrase denotes $Pr(HP|HP \vee F)$. Does it make sense for these to have different values? To do so involves distinguishing between today and tomorrow in a way that seems unwarranted by the setup of the problem. Shift invariance, it seems, is a requirement of rationality.

On the other hand, it is stipulated in the setup that coin tosses on distinct days are independent of each other. $Pr(H|P_n)$ is equal to $1/2$, for every n, no matter how large. The toss today is independent of every past toss; should it not also be independent of *all* the past tosses? Recall that all of these probabilities are meant to be predicated on the supposition that the setup is as described, which includes stipulation of independent tosses. For our credences, conditional on this setup, to violate independence, setting $Pr(H|P)$ equal to 1, seems no less irrational than violation of shift invariance.

4 Probabilities conditional on a σ-algebra

Consider, once again, Examples 1. As noted, an "obvious" answer to the question of the probability that a point chosen with uniform probability on the sphere lies in the Western hemisphere, conditional on the supposition that it lies on the equator, is $1/2$. For the question of the probability that the point lies closer to an equator than a pole, conditional on the supposition that it lies on the Greenwich meridian, both $1/2$ and $1/\sqrt{2}$ seem to have merit.

One way to think about question 1(a) is to imagine that, first, a circle of latitude is chosen, and then a point is chosen on that circle according a probability distribution conditional on the point lying on the circle. Because the area between two circles of latitude, at angles a, b, measured from the equator, is equal to

$$\frac{1}{2}\int_a^b \cos\phi \; d\phi,$$

the latitude Φ must be distributed according to

$$Pr(\Phi \in A) = \frac{1}{2}\int_A \cos\phi \; d\phi. \tag{14}$$

That is, Φ has density function

$$f_\Phi(\phi) = \frac{1}{2}\cos\phi. \tag{15}$$

The longitude Θ is distributed with uniform probability on $[-\pi, \pi]$, and so has density function

$$f_\Theta(\theta) = \frac{1}{2\pi}. \tag{16}$$

Latitude and longitude are independent random variables. That is,

$$Pr(\Phi \in A \,\&\, \Theta \in B)) = Pr(\Phi \in A)\; Pr(\Theta \in B) = \int_A f_\Phi(\phi)\; d\phi \int_B f_\Theta(\theta)\; d\theta. \qquad (17)$$

for all measurable $A \subseteq [-\pi/2, \pi/2]$ and $B \subseteq [-\pi, \pi]$.

What should the conditional distribution of the longitude Θ be taken to be, conditional on a given circle of latitude? We want the conditional probabilities to mesh properly with the unconditional probabilities. That is, we want to have

$$Pr(\Phi \in A \,\&\, \Theta \in B)) = \int_A Pr(\Theta \in B | \Phi \in \phi)\; f_\Phi(\phi)\; d\phi \qquad (18)$$

for all measurable A, B. The simplest way to do this, which is also the way that is naturally suggested by the independence of Θ and Φ, is to take $Pr(\Theta \in B | \Phi = \phi)$, for each B, to have the constant value $Pr(\Theta \in B)$, independent of ϕ. But it's not the only way. We can take any set of latitudes of measure zero, and choose distributions for Θ, conditional on $\Phi = \phi$ in that set, any way we want, and still satisfy the meshing condition (18). That means that (18) is compatible with *any* answer to question 1(a).

It is natural, however, to take $Pr(\Theta \in B | \Phi = \phi)$ to be, for each B, a continuous function of ϕ. This condition, together with the meshing condition (18), uniquely fixes

$$Pr(\Theta \in B \,|\, \Phi = \phi) = Pr(\Theta \in B). \qquad (19)$$

Similarly, we can define conditional distributions of latitude, conditional on meridian lines (lines of constant longitude), and demand that these also mesh with the unconditional probabilities:

$$Pr(\Phi \in A \,\&\, \Theta \in B) = \int_B Pr(\Phi \in A | \Theta = \theta)\; f_\Theta(\theta)\; d\theta \qquad (20)$$

This, together with the requirement that for each A, $Pr(\Phi \in A | \Theta = \theta)$ be a continuous function of θ, uniquely fixes

$$Pr(\Phi \in A | \Theta = \theta) = Pr(\Phi \in A), \qquad (21)$$

corresponding to conditional density functions

$$f_\Phi(\phi \,|\, \Theta = \theta) = \frac{1}{2} \cos\phi. \qquad (22)$$

Consider, now, question 1(b). We can imagine that a meridian is first chosen, and then a point chosen on that meridian. Using (21) yields the result that, conditional on any meridian, the probability is $1/\sqrt{2}$ that the chosen point is closer to the equator than to a pole.

On the other hand, since we are only imagining these things, we can also imagine the sphere partitioned by circles parallel to the circle containing the Greenwich meridian (see Figure 1) and imagine that first one of these circles is chosen, and then a point

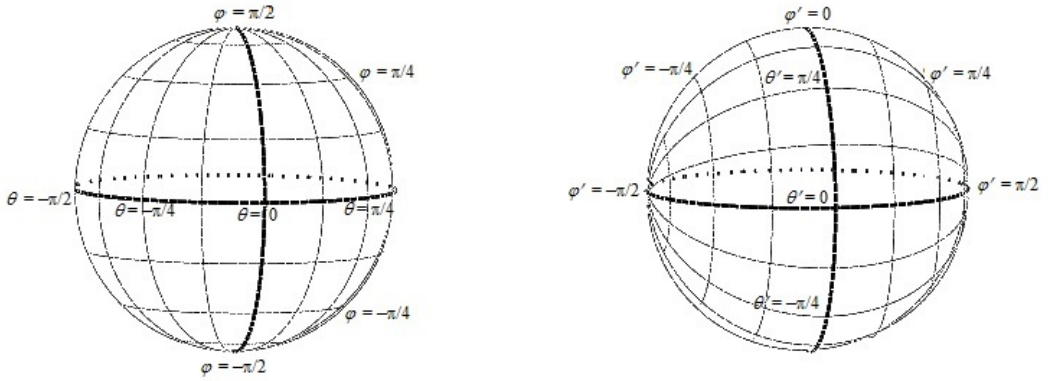

Figure 1: Two coordinatizations of the sphere.

chosen on that circle according to a probability distribution conditional on the circle. If this procedure is to yield uniform probabilities on the sphere, we must have the distributions on almost all of these circles be uniform, and this plus continuity militates a uniform distribution on all of them. This yields the answer 1/2 to question 1(b).

Which answer is correct? If the point on the sphere is, in fact, chosen according to one of the two-step procedures we have imagined, then such a set-up privileges one of the answers. But if the point is simply chosen, with uniform probability, on the sphere, then the set-up privileges neither answer, and, if one or the other has greater intuitive appeal, this may be because one is implicitly assuming one or the other scenario.

A circle is just a circle,[8] and the great circle containing the Greenwich meridian, *qua* circle on the sphere, is a element of many different partitions of the sphere. If we really think that $f_\Phi(\phi \,|\, \Theta = 0)$, as given by (22), is a conditional density function yielding the distribution of the random variable Φ conditional on the supposition that $\Theta = 0$, then it shouldn't matter how this supposition is described. The supposition can equally well be described using coordinates that take circles parallel to the great

[8]Oddly enough, this has been disputed. In connection with this example, E.T. Jaynes (2003, p. 470) writes,

> Nearly everybody feels that he knows perfectly well what a great circle is; so it is difficult to get people to see that the term 'great circle' is ambiguous until we specify what limiting operation is to produce it.

This strikes me as confused. One and the same great circle can be the limit of many different decreasing sequences of subsets of the sphere, but the circle is not itself *produced* by the limiting operation. Not so with probabilities conditional on a great circle, which, unless stipulated as primitive, are obtained via some limiting operation.

circle containing the Greenwich meridian as lines of latitude ϕ'. Then the great circle containing our original Greenwich meridian is the set of points for which $\phi' = 0$. On this circle the new longitude θ' differs from the old latitude ϕ by a constant, and, if we choose the zero-point of our new longitude as our old equator, we will have θ' equal to ϕ' on the circle. But a uniform distribution of the new longitude on circles of constant ϕ' requires a conditional density function

$$f_{\Theta'}(\theta' \,|\, \Phi' = \phi') = \frac{1}{2\pi} \tag{23}$$

It can't be the case that, conditional on the chosen point lying on the circle that is the great circle containing the Greenwich meridian of our first coordinatization and is the equator of our second, we have different conditional distributions depending on how we describe the circle. Taking (22) to yield the conditional distribution of Φ on this circle is incompatible with a uniform distribution of Θ', as given by (23).

The conclusion we should come to is that, though functions such as $Pr(\Phi \in A \,|\, \Theta = \theta)$ and $Pr(\Theta' \in B \,|\, \Phi' = \phi')$ are useful as calculational tools, *it is simply a mistake* to regard them as yielding conditional probabilities, conditional on point values of Θ and Φ', unless something in the set-up of the problem picks out one partition as privileged.

Quantities such as $Pr(\Phi \in A \,|\, \Theta = \theta)$, viewed as a function of θ, are instances of what is known as conditional probabilities with respect to a random variable. Let $\langle \Omega, \mathcal{A}, Pr \rangle$ be a probability space, and let X be a random variable, with distribution μ_X, and let $\sigma(X)$ be the σ-algebra generated by X. Let $f_A : \Omega \to \mathbb{R}$ be some $\sigma(X)$-measurable function.[9] (Note that this has the consequence that f_A is constant on sets of constant X). We will say that f_A is a conditional probability of A with respect to X iff, for all Borel sets Δ,

$$Pr(A \,\&\, X \in \Delta) = \int_{X \in \Delta} f_A(x)\, d\mu_X(x), \tag{24}$$

and write $f_A = P(A||X)$. These functions will be unique up to sets of probability zero; any two such functions are equal almost everywhere.

Conditional probabilities with respect to a random variable are special cases of conditional probabilities with respect to a σ-algebra. Let $\langle \Omega, \mathcal{A}, P \rangle$ be a probability space. For any σ-algebra $\mathcal{G} \subseteq \mathcal{A}$, and any $A \in \mathcal{A}$, a function $f_A : \Omega \to \mathbb{R}$ is a conditional probability of A with respect to $\mathcal{G}$ iff it is a $\mathcal{G}$-measurable function such that

$$P(AG) = \int_G f_A \, dP \tag{25}$$

for all $G \in \mathcal{G}$. We will write $f_A = P(A||\mathcal{G})$. (25) then becomes

$$P(AG) = \int_G P(A||\mathcal{G}) \, dP. \tag{26}$$

[9]This means: for any measurable subset Δ of $\mathbb{R}$, the set $f_A^{-1}(\Delta) = \{\omega \mid f_A(\omega) \in \Delta\}$ is in $\sigma(X)$.

Hájek (2003, 291) calls this "Kolmogorov's elaboration of the ratio formula." The existence of such functions is guaranteed by the Radon-Nikodym theorem (see, *e.g.*, Billingsley (2012, §32–33).)

Conditional probabilities, as usually conceived, are a special case of conditional probabilities with respect to a σ-algebra. Let $\{G_i\}$ be a countable partition, and let $\mathcal{G}$ be the σ-algebra generated by this partition. Since the elements of the partition $\{G_i\}$ are atoms of this σ-algebra, and $P(A||\mathcal{G})$ is required to be a $\mathcal{G}$-measurable function, it must be a constant function on each G_i. Let $P(A|G_i)$ be the value that $P(A||\mathcal{G})(\omega)$ takes on for $\omega \in G_i$. Then the condition that (26) hold for all $G \in \mathcal{G}$ is equivalent to the condition that

$$P(AG_i) = P(A|G_i)P(G_i), \tag{27}$$

which, of course, yields the familiar ratio formula for $P(A|G_i)$ whenever $P(G_i) > 0$. In this sense, we have a generalization of conditional probabilities.

When an agent learns which element of $\{G_i\}$ is true, she at the same time learns the truth value of each proposition in the σ-algebra $\mathcal{G}$. The heuristic idea behind the introduction of probabilities conditional on more general σ-algebras is to mimic this. A random variable X partitions the space Ω of events into sets of constant X. To learn the value of X is to learn which of these sets ω is in, and thereby learn, for every set $\Delta \in \sigma(X)$, whether or not $\omega \in \Delta$.

Let $\mathcal{G} \subseteq \mathcal{A}$ be a σ-algebra that contains atoms—that is, elements of $\mathcal{G}$ with no non-empty proper subsets in $\mathcal{G}$—that cover Ω. For $A \in \mathcal{A}$, let $P(A||\mathcal{G})$ be a conditional probability of A with respect to $\mathcal{G}$. If G is an atom of $\mathcal{G}$, then $P(A||\mathcal{G})$ must take on a constant value on G. Should we regard this value, the value of $P(A||\mathcal{G})$ for $\omega \in G$, as the probability of A conditional on the proposition G?

There are two sorts of problems with this. The first is technical and local, in that it applies only to certain σ-algebras that we might dismiss as pathological. Nonetheless, it should give us pause, as it shows that the heuristic motivation of the characterization of $P(A||\mathcal{G})$, namely, as conditional probabilities resulting from an experiment in which it is learned, for each element G of a σ-algebra $\mathcal{G}$, whether or not $\omega \in \mathcal{G}$, can break down. The second sort of problem is conceptual and global, and poses a serious objection to taking the value of $P(A||\mathcal{G})$ for $\omega \in G$, as the probability of A conditional on the proposition G (except in special circumstances, to be discussed in the next section).

The first problem is this. On the heuristic view that $P(A||\mathcal{G})$, evaluated on some atom G of $\mathcal{G}$, yields the probability of A appropriate to learning that $\omega \in G$, we would expect that, if $\mathcal{G}$ is a σ-algebra whose atoms are all the singleton sets, then $P(A||\mathcal{G})$ would be equal to 1 if ω is in A and 0 if not, since learning which atom of $\mathcal{G}$ obtains is complete information about ω. But this won't always be the case. Let our probability space be the unit interval with Lebesgue measure. Let $\mathcal{G}$ consist of the the smallest σ-algebra containing all of the singleton sets; this consists of the countable sets and their complements. Now let A be any set with $P(A) \in (0, 1)$. It is easy to see that $P(A||\mathcal{G})(\omega)$ must be equal to $P(A)$ for almost all ω, violating our expectation that it will everywhere be equal to 0 or 1.[10]

[10]This is example 33.11 of Billingsley (2012).

The second problem is the one we have already seen in connection with the Borel-Kolmogorov paradox, and it is more serious. Let G be an atom of a σ-algebra $\mathcal{G}$. Though, for any G with $P(G) = 0$, the condition (26) leaves the value of $P(A||\mathcal{G})$ on G undetermined, the condition together with other natural constraints, such as requiring $P(A||\mathcal{G})$ to be a continuous function, can, as we have seen, determine the value of $P(A||\mathcal{G})$ on G. But this is not enough to warrant taking this value as the probability of A, conditional on G, as the same set G will be an atom of other σ-algebras, and the same considerations might dictate that, for some other σ-algebra $\mathcal{G}'$ containing G, the value that $P(A||\mathcal{G}')$ has on G be different from the value that $P(A||\mathcal{G})$ on G. In cases, such as the sphere example, in which the set-up privileges neither σ-algebra, it would be a mistake to take either of these values (or any other) as *the* probability of A conditional on G.

This is, of course, pretty much the standard view. Kolmogorov, in his discussion of the Borel paradox, writes, "This shows that the concept of a probability conditional on an isolated given hypothesis whose probability equals 0 is inadmissible" (Kolmogorov, 1950, p. 51). Taking up this suggestion, Easwaran concludes,

> this means we must view conditional probability as (in general) a three-place function, depending not only on A and G, but also the partition $\mathbf{G}$ defining the set of "relevant alternatives" to G. In particular cases, this partition will be specified by the experiment an agent is considering G as an outcome to, or the set of alternative hypotheses under consideration, or some other contextual factor. Thus, we must think of conditional degree of belief as a function $P(A|G, \mathbf{G})$ rather than just $P(A|G)$ (Easwaran, 2011, pp. 143–44).

We should ask: under what conditions will there be a set of relevant alternatives that is uniquely picked out by the set-up?

It is frequently suggested, as in the quotation from Easwaran, that it is the experiment that yields the data that determines a relevant partition (see also the discussion in Rényi 2007a, §2.1). On this rationale, though, it is hard to see that we would ever need to go beyond a finite partition. Unless we are entertaining the fiction of agents with infinite powers of discrimination, there are only finitely many distinguishable alternatives as to the outcome of any experiment. (This is even easier to see in these days in which laboratory equipment has digital readout than it was in the old days of pointers and dials!) Even if we do imagine agents with infinite powers of discrimination, the set of alternatives they could record, using a finite alphabet, in a lab notebook of finite capacity, is a finite set.

Unproblematic null-condition conditional probabilities are not as commonplace as some of the literature might suggest. However, there are cases in which the set-up of a problem *does* permit one to speak unambiguously of the the probability of an event conditional on a null proposition. In those cases, null-condition conditional probabilities are unobjectionable, and they can be useful, though they are not indispensable.

5 Unproblematic Null-Condition Probabilities

5.1 Likelihood functions

It is common, in statistical practice, to regard outcomes of some experiment as being generated by an incompletely known probability distribution characteristic of the experimental set-up; data gathered is used to gain information about that distribution. We commonly consider a family of candidate distributions; typically this family is indexed by some set of parameters. For instance, we might regard an experimentally measurable variable as being normally distributed with unknown mean μ and unknown variance σ^2. A data-set is generated, and is used to gain information about the values of the parameters.

Let Ω be the set of possible outcomes of an experiment, let $\mathcal{F}$ be the set of measurable subsets of Ω, and, for every value of θ in some parameter-space Γ, let P_θ be a probability distribution on $\langle \Omega, \mathcal{F} \rangle$. For fixed $E \in \mathcal{F}$, the function $\mathcal{L}_E(\theta) = P_\theta(E)$, considered as a function of θ, is called a *likelihood function*.

In standard, frequentist statistics, the parameter space is not itself subject to probabilistic considerations; it is regarded as nonsensical to ascribe probabilities, prior or posterior, to propositions regarding values of the parameters. Hence, P_θ is not regarded as a conditional probability distribution, conditional on a proposition of probability 0.

On a Bayesian approach, on the other hand, one also ascribes probabilities to propositions regarding the values of the parameters, and the process of gaining information about the parameter values is modelled by conditionalization on the experimental result. Let $\mathcal{G}$ be a σ-algebra of subsets of Γ. Let $\mathcal{H}$ be the smallest σ-algebra containing $\mathcal{F} \times \mathcal{G}$. Suppose that, for each $E \in \mathcal{F}$, the likelihood function $\mathcal{L}_E(\theta)$ is a $\mathcal{G}$-measurable function. Then, given a probability measure Q on $\langle \Gamma, \mathcal{G} \rangle$, we can form a new probability space whose event space is the Cartesian product $\Omega \times \Gamma$ of the experimental outcome space and the parameter space, and whose measurable sets are the set $\mathcal{H}$: we define a probability measure Pr as the unique countably additive extension to $\mathcal{H}$ of the function that, on $\mathcal{F} \times \mathcal{G}$, is given by

$$Pr(F \times G) = \int_G \mathcal{L}_F(\theta)\, dQ. \tag{28}$$

We now have a probability space $\langle \Omega \times \Gamma, \mathcal{H}, Pr \rangle$. The experimental outcome X, and parameter value Θ, are random variables on this probability space. The σ-algebra $\sigma(X)$ that consists of propositions about the experimental outcome is $\mathcal{F} \times \Gamma$, and $\sigma(\Theta)$, the σ-algebra that consists of propositions about parameter values, is $\Omega \times \mathcal{G}$. One can readily verify that a version of conditional probability with respect to $\sigma(\Theta)$ is obtained by setting

$$P(E||\Theta)(\omega) = \mathcal{L}_E(\Theta(\omega)). \tag{29}$$

Any version of conditional probability with respect to $\sigma(\Theta)$ will have to agree with (29) with probability one.

Is it permissible to regard the values that $P(E||\Theta)$ takes on on the atoms of $\sigma(\Theta)$ as probabilities conditional on null propositions? There is a natural one-one correspondence between the atoms of $\sigma(\Theta)$ and the points in the parameter space Γ. In this case, we have a σ-algebra that is picked out as special by the set-up of the problem; the random variable Θ represents the parameters of the system about which we are trying to gain information, and the atoms of the σ-algebra $\sigma(\Theta)$ correspond to maximal specification of these parameters. In this case, there seems no threat of ambiguity due to variant choices of σ-algebra to conditionalize on, and we can, relatively unproblematically, regard these values as null-condition probabilities.

We can form a conditional probability space by taking the set $\mathcal{B}$ of permissible conditions to include, in addition to all propositions with positive probability, also the atoms of $\sigma(\Theta)$, corresponding to point values of our parameters. This still leaves us with a set of conditions that, though it goes beyond the standard set, is still fairly sparse compared to the full set $\mathcal{H}$ of measurable sets.

It is not uncommon to deal with nested families of models, in which the parameter space of one model is a lower-dimensional subspace of the parameter space of another. This might come about, for example, by considering a model in which the value of some parameter is fixed, or two parameters are constrained to be equal. We will want to retain the same likelihoods in the reduced model. We will also want a probability distribution over the the reduced space. Here again the issue illustrated by the Borel-Kolmogorov paradox resurfaces; the probability distribution on the higher-dimensional space does not determine a distribution on the lower-dimensional space, and defining one via a limiting procedure will lead to differing results, depending on the procedure chosen. It is a mistake to regard the lower-dimensional model as being obtained, in a straightforward way, from the higher-dimensional model via conditionalization.

Though, in this case, it is permissible to treat the likelihoods as probabilities conditional on point-values of the parameters, it is by no means necessary to do so. All of our standard statistical reasoning goes through if we restrict our domain of conditionalization to the traditional choice of sets with positive probability.

If we obtain evidence E about the experimental outcome, then we can update our credences about parameter values via conditionalization. For any measurable subset Δ of parameter-space,

$$Pr(\Theta \in \Delta) \rightarrow Pr_E(\theta \in \Delta) = \frac{Pr(E \,\&\, \Theta \in \Delta)}{Pr(E)}. \tag{30}$$

If the prior distribution of Θ is given by a density function μ, then the process of conditionalization yields a new density function μ_E. In order for (30) to be satisfied, we must have, for almost all θ,

$$\mu_E(\theta) = \frac{\mathcal{L}_E(\theta)\,\mu(\theta)}{Pr(E)}. \tag{31}$$

This is often called the *continuous form of Bayes' theorem.* Thinking of (31) as a form of Bayes' theorem invites to think of $\mathcal{L}_E(\theta)$ as the probability of E conditional on a

point value of the parameter θ. But the use of the new density μ_E is to generate new probabilities $Pr_E(\Theta \in \Delta)$, and this can be done via (30), and there is no need to invoke probabilities conditional on null subsets of the parameter space.

All that we need for Bayesian statistical inference is the probability space $\langle \Omega, \mathcal{H}, Pr \rangle$, and operations on this, including conditionalization on new evidence, can go through in the standard way, without invoking any conditional probabilities conditional on null subsets of the parameter space. We can, if convenient, work with the likelihood functions $\mathcal{L}_E(\theta)$, whose existence is guaranteed by the Radon-Nikodym theorem. But there is no *need* to regard these as *bona fide* conditional probabilities, and their usefulness as calculational tools does not depend on any such interpretation.

5.2 Stochastic processes

In the theory of stochastic processes, we deal with a set $\{X_t \,|\, t \in T\}$ of random variables, where the index t is to be thought of as a time index (which may be continuous or discrete). As an example, consider the following simple two-step process, adapted from Bayes (1763). A ball is thrown onto a square table $ABCD$, with unit sides, with uniform probability on the square for its landing place. A line drawn through its landing point, parallel to AD. We then throw a second ball, again with uniform probability, and are provided with a report of whether the second ball landed to the left or the right of the line we drew. In this case, it is unproblematic to say that, conditional on the first ball's landing at a distance x from the left side of the table, the chance of the second ball landing to the left of the line is x.

But we don't have to; everything we need to say about the process can be said without invocation of null-condition conditional probabilities. Let X_1 be the random variable that represents the distance of the landing place of the first ball from the left side of the table, and let X_2 be the random variable that takes on the value L or R depending on whether the second ball lands to the left or right of the line through the landing-place of the first ball. We can specify the two-step process by saying that X_1 is uniformly distributed on $(0, 1)$, and that conditional probabilities for X_2 are given by

$$\begin{array}{l} Pr(X_2 = L \,|\, X_1 = x) = x \\ Pr(X_2 = R \,|\, X_1 = x) = 1 - x. \end{array} \tag{32}$$

but we can also achieve the same effect by saying that joint probabilities regarding X_1 and X_2 satisfy

$$\begin{aligned} Pr(X_2 = L \,\&\, X_1 \in \Delta) &= \int_\Delta x \, dx \\ Pr(X_2 = R \,\&\, X_1 \in \Delta) &= \int_\Delta (1 - x) \, dx \end{aligned} \tag{33}$$

for every Borel set $\Delta \subseteq (0, 1)$. Null-condition conditional probabilities, though they may provide a useful way of talking, are not needed to specify the stochastic process.

More generally, given a stochastic process involving random variables $\{X_t \,|\, t \in T\}$, for any time t_0 we can consider the set of random variables with $t \leq t_0$, and form a

σ-algebra $\mathcal{T}_0$ generated by this set of random variables. For any proposition of the form $X_r \in \Delta$, we will have conditional probabilities with respect to $\mathcal{T}_0$, $P(X_r \in \Delta || \mathcal{T}_0)$. The values these take on the atoms of $\mathcal{T}_0$ may be regarded as probabilities conditional on a full specification of events up to t_0, even if these atoms have zero probability.

Cautions that by now are familiar are in place: though the set-up gives us a privileged σ-algebra, namely, the σ-algebra corresponding to a full specification of events up to t_0, including these events in our set of admissible conditions for conditional probability still leaves us with a rather spare set of conditions, and problems and ambiguities may arise if we seek to include in our set of conditions null propositions with less than complete information about the past. Secondly, the stochastic process only specifies these conditional probabilities for almost all histories; different versions of the conditional probabilities may differ on probabilities conditional on past histories in some set of measure zero. These are not taken as corresponding distinct stochastic processes, as they yield the same probability for any set of events.

The Eternal Coin example of §3.3 illustrates this latter point. Let $\mathcal{P}$ be the σ-algebra consisting of propositions about results of coin tosses to the past of today. The atoms of this σ-algebra comprise all possible complete specifications of the past; the proposition P, that the coin landed heads every day in the past, is one such. The proposition H, that the coin lands heads today, is independent of the σ-algebra $\mathcal{P}$. That is,

$$Pr(AH) = Pr(A)\, Pr(H) \tag{34}$$

for all $A \in \mathcal{P}$. This entails that we must have

$$P(H||\mathcal{P})(u) = Pr(H) = 1/2 \tag{35}$$

for almost all u in our event space. But this doesn't preclude Dorr, or anyone else so inclined, from assigning the value 1, or any other value, to the probability of H conditional on the proposition P, or on any set of propositions comprising a set of measure zero. Distinct choices of this sort yield the same probabilities for all propositions.

If it is a physical process that we are modelling, this, arguably, is all that matters. Suppose that we are formulating a physical theory with stochastic dynamics, and formulate the theory in terms of transition probabilities, that is, probabilities about future events conditional on past events. Two formulations that agree on transition probabilities for all but a set of histories of measure zero attribute the same probabilities to all sets of events. On any reasonable criterion of individuation of physical theories, these should count as variant formulations of the same theory. This means that, if we want to think of the laws of a stochastic physical theory as specifying, for every complete history up to time t, conditional probabilities concerning events to the future of t, what the laws actually specify is an *equivalence class* of such conditional probabilities, where two sets of transition probabilities are equivalent if they agree on almost all histories.

5.3 The Principal Principle

The Principal Principle, so named by Lewis (1980), is the prescription that your credence at time t in a proposition A, conditional on the supposition that the chance at t of A is x and any admissible proposition, be x. That is,

$$Cr_t(A \mid ch_t(A) = x \,\&\, E) = x. \tag{36}$$

for any admissible E, where "[a]dmissible propositions are the sort of information whose impact on credence about outcomes comes entirely by way of credence about the chances of those outcomes" (Lewis, 1980, p. 272). This is to be true for every value of x in $[0, 1]$. Any credences about the chance of A will assign zero credence to uncountably many singleton sets. Thus, it looks as if the Principal Principle *commits* us to conditionalizing on null propositions.

This, again, is unobjectionable, as we have a distinguished σ-algebra, consisting of propositions of the form $ch_t(A) \in \Delta$, where Δ ranges over Borel subsets of $[0, 1]$. But use of the Principal Principle itself does not by itself commit us to null-condition probabilities. The essential content of the Principle can be expressed as the prescription that, for every interval $\Delta \subseteq [0, 1]$ with $Cr_t(ch_t(A) \in \Delta) > 0$, and any admissible E,

$$Cr_t(A \mid E \,\&\, ch_t(A) \in \Delta) \in \Delta. \tag{37}$$

Thus, even without primitive probabilities conditional on null propositions, we get what we need.

This readily extends to credences about multiple propositions. For any finite set $\mathbf{A} = \{A_1, ..., A_n\}$ of propositions, we require that, for all measurable $\Delta \subseteq [0, 1]^n$, with $Cr_t(ch_t(\mathbf{A} \in \Delta)) > 0$,

$$Cr_t(\cap_i A_i \mid E \,\&\, ch_t(\mathbf{A}) \in \Delta) \in \mathrm{Conv}(\Delta), \tag{38}$$

where $\mathrm{Conv}(\Delta)$ is the convex hull of Δ.

6 Conclusion

Talk of probabilities conditional on zero-probability propositions is common in the philosophical literature. There is nothing *necessarily* incoherent in such talk, and we may, for certain purposes, find it convenient to include such propositions in the stock of proposition on which we conditionalize. But the motivations for doing so have been exaggerated.

Moreover, though symmetry considerations may guide us in choice of probability distribution conditional on null propositions, such considerations can be less than reliable guides. Imposing the requirement that the conditional probability space be invariant under all symmetries of the unconditional probability space is excessively restrictive. If we want to extend our conditional probability space to include conditionalization on null propositions, we will have to be selective about which symmetries

of the unconditional probability space we impose on the conditional probability space. In some cases—such as the sphere—there may be a natural choice of which symmetries to impose. In other cases, of which Dorr's Eternal Coin is a striking example, symmetry considerations will lead us in opposing directions, without a clear choice to be made.

If, nonetheless, you want to include null proposition in your set of conditions: proceed with caution, and with care to state explicitly how your conditional probability space is to be constructed.

7 Acknowledgments

I thank Alan Hájek, Bill Harper, and Joshua Luczak for helpful discussions on these matters. This work was sponsored, in part, by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC).

Appendix 1 Terminology

1.1 Probability Spaces

For any set S, an *algebra* of subsets of S is a set of subsets of S that contains S and is closed under complementation and unions. A *σ-algebra* of subsets of S is an algebra that is closed under countable unions. For the real line $\mathbb{R}$, we define the Borel sets as the smallest σ-algebra containing all open intervals.

If $\mathcal{A}$ is an algebra of subsets of S, a function $P : \mathcal{A} \to \mathbb{R}$ is *additive* iff, for any disjoint $A, B \in \mathcal{A}$,

$$P(A \cup B) = P(A) + P(B).$$

If $\mathcal{A}$ is a σ-algebra of subsets of S, a function $P : \mathcal{A} \to \mathbb{R}$ is *countably additive* iff, for any sequence $\{A_i\}$ of disjoint sets in $\mathcal{A}$,

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

A *probability space* is a triple $\langle S, \mathcal{A}, Pr \rangle$, where S is a set, to be thought of as the set of elementary events, $\mathcal{A}$ is an algebra of subsets of S, which are the sets of events (propositions) to which probabilities will be ascribed, and $Pr : \mathcal{A} \to \mathbb{R}$ is a probability function, that is, a positive, additive set function with $Pr(S) = 1$. Since we will have reasons to consider probability functions that are not countably additive, we depart from tradition in not assuming countable additivity unless explicitly stated. If we require countable additivity, then $\mathcal{A}$ is required to be a σ-algebra, and we will refer to Pr as a *probability measure*.

If $\langle S, \mathcal{A}, Pr \rangle$ is a probability space, a *random variable* is a measurable function $X : S \to \mathbb{R}$, that is, a function such that, for any Borel set B, the set

$$X^{-1}(B) = \{\omega \in S \mid X(\omega) \in B\}$$

is in $\mathcal{A}$. A random variable X generates a subalgebra of $\mathcal{A}$, called $\sigma(X)$, which is the set of all $X^{-1}(B)$, as B ranges over Borel subsets of the real line.

1.2 Conditional Probability Spaces

Following Rényi (1955, 2007a,b),[11] we define a *conditional probability space* as a quadruplet $\langle S, \mathcal{A}, \mathcal{B}, P \rangle$, where S is a set of events, $\mathcal{A}$ an algebra of subsets of S, $\mathcal{B}$ a subset of $\mathcal{A}$, to be thought of as the set of events on which we may conditionalize, and P is

[11] Though inspired and instructed by Rényi's treatment, this definition departs from Rényi in two ways. First, Rényi requires $P_B(A)$ to be defined for every $A \in \mathcal{A}$. This may be undesirable; see Appendix 2. Second, Rényi requires countable additivity, and we leave open the possibility of conditional probability functions that are merely finitely additive. Rényi (2007a, §2.2) adds the further conditions that the set $\mathcal{B}$ be closed under finite disjunctions, and that it contain a sequence $\{B_n\}$ that covers Ω. A subset of a σ-algebra $\mathcal{A}$ satisfying these two conditions, and not containing the null set, Rényi calls a *bunch* of sets.

a function that takes $B \in \mathcal{B}$ to a function $P_B : \mathcal{A}_B \to \mathbb{R}$, where, for each B, $\mathcal{A}_B$ is a subalgebra of $\mathcal{A}$, and

i). For each $B \in \mathcal{B}$
   (a) $P_B(A) \geq 0$ for all $A \in \mathcal{A}_B$.
   (b) For all $A \in \mathcal{A}$, if $B \subseteq A$, then $A \in \mathcal{A}_B$ and $P_B(A) = 1$.
   (c) For disjoint $A, A' \in \mathcal{A}_B$, $P_B(A \vee A') = P_B(A) + P_B(A')$.

ii). For all $B, C \in \mathcal{B}$ and $A, B \in \mathcal{A}_C$, if $BC \in \mathcal{B}$ then

$$P_C(AB) = P_{BC}(A)\, P_C(B),$$

provided that $C \in \mathcal{A}_B$ and $A \in \mathcal{A}_{BC}$.

A conditional probability space can be thought of as a family of probability spaces $\{\langle S, \mathcal{A}_B, P_B\rangle \mid B \in \mathcal{B}\}$, required to mesh with each other via (ii).

It is an immediate consequence of (ii) that, for any $C \in \mathcal{B}$ and $B \subseteq C$, if $B \in \mathcal{A}_C$ and $P_C(B) > 0$, then, for all $A \in \mathcal{A}_C$,

$$P_B(A) = \frac{P_C(AB)}{P_C(B)} \tag{39}$$

provided that $B \in \mathcal{B}$ and $A \in \mathcal{A}_B$. This allows us to define probabilities conditional on B, provided they don't clash with those yielded by some other $D \in \mathcal{B}$ such that $B \subseteq D$, $B \in \mathcal{A}_D$, and $P_D(B) > 0$. For this reason, we will usually assume the further condition,

iii). For all $C \in \mathcal{B}$ and $B \subseteq C$, if $B \in \mathcal{A}_C$ and $P_C(B) > 0$, then $B \in \mathcal{B}$ and $\mathcal{A}_C \subseteq \mathcal{A}_B$.

Given a probability space $\langle S, \mathcal{A}, Pr\rangle$, let $\mathcal{A}^*$ be the subset of $\mathcal{A}$ consisting of sets B with $Pr(B) > 0$. Let P^* be the function that maps $B \in \mathcal{A}^*$ to the probability function $P_B : \mathcal{A} \to [0, 1]$, given by

$$P_B(A) = \frac{Pr(AB)}{Pr(B)}. \tag{40}$$

Then $\langle S, \mathcal{A}, \mathcal{A}^*, P^*\rangle$ is a conditional probability space, corresponding to the standard choice of having conditional probability defined only when the condition has nonzero probability.

We will say that a probability space $\langle S, \mathcal{A}, Pr\rangle$ is invariant under a bijection $\mathsf{T} : S \to S$ if and only if $\mathsf{T}(\mathcal{A}) = \mathcal{A}$ and, for all $A \in \mathcal{A}$, $Pr(\mathsf{T}(A)) = Pr(A)$. Similarly, a conditional probability space $\langle S, \mathcal{A}, \mathcal{B}, P\rangle$ is invariant under T if and only if $\mathsf{T}(\mathcal{A}) = \mathcal{A}$, $\mathsf{T}(\mathcal{B}) = \mathcal{B}$, and, for all $B \in \mathcal{B}$, $\mathcal{A}_{\mathsf{T}(B)} = \mathsf{T}(\mathcal{A}_B)$ and $P_{\mathsf{T}(B)}(\mathsf{T}(A)) = P_B(A)$ for all $A \in \mathcal{A}_B$.

Appendix 2 A rotationally invariant conditional probability space

Let S be the set of points on the unit sphere, let λ_S be uniform measure on the unit sphere, and let $\mathcal{L}_S$ be the set of all λ_S-measurable subsets of S. Let $\mathcal{C}$ be the set of all

circles on S. For each circle $C \in \mathcal{C}$, let λ_C be uniform measure on the circle C, and let $\mathcal{L}_C$ be the set of λ_C-measurable subsets of C.

We can construct standard conditional probability spaces $\langle S, \mathcal{L}_S, \mathcal{L}_S^*, P_S^* \rangle$ and $\langle C, \mathcal{L}_C, \mathcal{L}_C^*, P_C^* \rangle$, which include conditionalization only on sets of nonzero measure. We want to extend $\langle S, \mathcal{L}_S, \mathcal{L}_S^*, P_S^* \rangle$ to include, at minimum, conditionalization on circles, in such a way that probabilities conditional on these circles yield uniform probabilities on the circles.

Let us take $\mathcal{B}$ to be

$$\mathcal{B} = \mathcal{L}_S^* \cup \bigcup_{C \in \mathcal{C}} \mathcal{L}_C^*. \tag{41}$$

Note that each element of $\mathcal{B}$ is either in $\mathcal{L}_S^*$ or is a subset of a *unique* circle C. Take $\mathcal{A}_B$ to be $\mathcal{L}_S$ for $B \in \mathcal{L}_S^*$. For $B \in \mathcal{L}_C^*$, take $A \in \mathcal{A}_B$ iff $A' \subset A$ for some $A' \in \mathcal{L}_C$. Define

$$P_B(A) = \begin{cases} \frac{\lambda_S(AB)}{\lambda_S(B)}, & B \in \mathcal{L}_S^*; \\ \\ \frac{\lambda_C(AB)}{\lambda_C(B)}, & B \in \mathcal{L}_C^*. \end{cases} \tag{42}$$

We have constructed a conditional probability space that is invariant under all rigid rotations, and includes conditionalization on circles and some subsets of circles.

For any $A \in \mathcal{L}_S$ with $\lambda_S(A) = 0$, every subset of A is a measurable set, and is assigned measure 0. Since each circle C has $\lambda_S(C) = 0$, this means that every subset of C is in $\mathcal{L}_S$. Since *not* every subset of C is in $\mathcal{L}_C$, $P_C(A)$ is not defined for arbitrary $A \in \mathcal{L}_S$. We might want to extend P_C so that it is defined on all $A \in \mathcal{L}_S$. But, as already mentioned, we can do so, and preserve rotational invariance, only at the price of sacrificing countable additivity. We can't get all that we want.

Appendix 3 The Eternal Coin: Proof of Dorr's theorem

We will speak in general terms but readers should think of the example at hand, that of the Eternal Coin. We assume Axiom (iii) of Appendix 1.2.

Suppose there is a proposition P, and a transformation T, such that $\mathsf{T}(P) \models P$. If $Pr(P) > 0$, then

$$Pr(\mathsf{T}(P) \mid P) = \frac{Pr(\mathsf{T}(P))}{Pr(P)}, \tag{43}$$

and so T-invariance would entail that $Pr(\mathsf{T}(P) \mid P) = 1$. Furthermore, if there exists a proposition Z such that $\mathsf{T}(Z) = Z$, $P \models Z$, and $Pr(P|Z) > 0$, then

$$Pr(\mathsf{T}(P) \mid P) = \frac{Pr(\mathsf{T}(P) \mid Z)}{Pr(P \mid Z)}, \tag{44}$$

and so, once again, T-invariance would entail that $Pr(\mathsf{T}(P) \mid P) = 1$.

But$\mathsf{T}$-invariant propositions of the right sort may be hard to come by, and there may be no such Z. Suppose, however, that there exist propositions X, Z, such that

$P \models X \models Z$, and $\mathsf{T}(P) \models \mathsf{T}(X) \models Z$. Then, if P, $\mathsf{T}(P)$, X, and $\mathsf{T}(X)$ are all in $\mathcal{A}_Z$, and $Pr(P|Z) > 0$, and if $Pr(\mathsf{T}(P)\,|\mathsf{T}(X))$ is defined, we have

$$\begin{aligned} Pr(P \,|\, Z) &= Pr(P \,|\, X)\; Pr(X \,|\, Z); \\ Pr(\mathsf{T}(P) \,|\, Z) &= Pr(\mathsf{T}(P)\,|\mathsf{T}(X))\; Pr(\mathsf{T}(X) \,|\, Z), \end{aligned} \tag{45}$$

and so,

$$Pr(\mathsf{T}(P) \,|\, P) = \frac{Pr(\mathsf{T}(P) \,|\, \mathsf{T}(X))}{Pr(P \,|\, X)}\; \frac{Pr(\mathsf{T}(X) \,|\, Z)}{Pr(X \,|\, Z)}. \tag{46}$$

Now suppose that there is also a proposition F such that $\mathsf{T}^{-1}(F) \models F$, with $Pr(F|Z) > 0$. Suppose, also, that $\mathsf{T}^{-1}(F) \models X$ and $F \models \mathsf{T}(X)$. Then

$$Pr(\mathsf{T}^{-1}(F) \,|\, F) = \frac{Pr(\mathsf{T}^{-1}(F) \,|\, X)}{Pr(F \,|\, \mathsf{T}(X))}\; \frac{Pr(X \,|\, Z)}{Pr(\mathsf{T}(X) \,|\, Z)}. \tag{47}$$

Multiplying (46) and (47) gives us,

$$Pr(\mathsf{T}(P) \,|\, P)\; Pr(\mathsf{T}^{-1}(F) \,|\, F) = \frac{Pr(\mathsf{T}(P) \,|\, \mathsf{T}(X))}{Pr(P \,|\, X)}\; \frac{Pr(\mathsf{T}^{-1}(F) \,|\, X)}{Pr(F \,|\, \mathsf{T}(X)}. \tag{48}$$

So far, we haven't invoked any symmetry assumptions. If we impose T-invariance, we have

$$\begin{aligned} Pr(\mathsf{T}(P) \,|\, \mathsf{T}(X)) &= Pr(P \,|\, X); \\ Pr(\mathsf{T}^{-1}(F) \,|\, X) &= Pr(F \,|\, \mathsf{T}(X)), \end{aligned} \tag{49}$$

and (48) becomes

$$Pr(\mathsf{T}(P) \,|\, P)\; Pr(\mathsf{T}^{-1}(F) \,|\, F) = 1, \tag{50}$$

from which it follows that

$$Pr(\mathsf{T}(P) \,|\, P) = Pr(\mathsf{T}^{-1}(F) \,|\, F) = 1. \tag{51}$$

Now, since we have assumed that $\mathsf{T}(P) \models P$ and $\mathsf{T}^{-1}(F) \models F$, there always do exists Z, X satisfying the conditions stipulated. Take Z to be $P \vee F$, and take X to be $P \vee \mathsf{T}^{-1}(F)$. Then $\mathsf{T}(X)$ is $\mathsf{T}(P) \vee F$.

To sum up: we have established

Proposition 1 *Let $\langle \Omega, \mathcal{A}, \mathcal{B}, P\rangle$ be a conditional probability space satisfying condition (iii). Suppose there exist a transformation T of Ω and propositions P, F, such that $Z = P \vee F \in \mathcal{B}$ and $P, F \in \mathcal{A}_Z$, such that*

i). (a) $\mathsf{T}(P) \models P$;
(b) $\mathsf{T}^{-1}(F) \models F$;

ii). (a) $Pr(P|Z) > 0$;
(b) $Pr(F|Z) > 0$.

Then

$$Pr(\mathsf{T}(P)\,|\,P)\ Pr(\mathsf{T}^{-1}(F)\,|\,F) = \frac{Pr(\mathsf{T}(P)\,|\,\mathsf{T}(X))}{Pr(P\,|\,X)}\ \frac{Pr(\mathsf{T}^{-1}(F)\,|\,X)}{Pr(F\,|\,\mathsf{T}(X))},$$

where $X = P \vee \mathsf{T}^{-1}(F)$.

If, further,

iii). (a) $Pr(\mathsf{T}(P)\,|\,\mathsf{T}(X)) = Pr(P\,|\,X)$*;*

(b) $Pr(\mathsf{T}^{-1}(F)\,|\,X) = Pr(F\,|\,\mathsf{T}(X))$*;*

then

$$Pr(\mathsf{T}(P)\,|\,P) = Pr(\mathsf{T}^{-1}(F)\,|\,F) = 1.$$

Applied to the Eternal Coin, let T be S_1, which shifts everything forward one day. P, as before, is the proposition that the coin handed heads every day in the past, and F, the proposition that the coin will land heads every day in the future. Let H be the proposition that the coin lands heads today. Then $\mathsf{S}_1(P)$ is HP, and $\mathsf{S}_1^{-1}(F)$ is HF. Clearly, $HP \models P$ and $HF \models F$. If

$$\begin{array}{l} Pr(P|P \vee F) > 0; \\ Pr(F|P \vee F) > 0; \end{array} \tag{52}$$

and if

$$\begin{array}{l} Pr(HP|HP \vee F) = Pr(P|P \vee HF); \\ Pr(F|HP \vee F) = Pr(HF|P \vee HF), \end{array} \tag{53}$$

then

$$Pr(H|P) = Pr(H|F) = 1. \tag{54}$$

We can run the same argument with S_k, for any positive k, yielding the conclusion that, for every $n \geq 0$, the probability conditional on P that the coin lands heads today and n days into the future is 1, as is the probability, conditional on F, that the coin lands heads today and n days into the past.

References

Banach, S. (1923). Sur le problème de la mesure. *Fundamenta Mathematicae 4*, 7–33.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions 53*, 370–418.

Billingsley, P. (2012). *Probability and Measure, Anniversary Edition*. Hoboken, NJ: Wiley.

Dorr, C. (2010). The Eternal Coin: A puzzle about self-locating conditional credence. *Philosophical Perspectives 24*, 189–205.

Easwaran, K. (2011). The varieties of conditional probability. In P. Bandyopadhyay and M. Forster (Eds.), *Handbook of the Philosophy of Science. Philosophy of Statistics*, pp. 137–148. Amsterdam: North-Holland.

Hájek, A. (2003). What conditional probability could not be. *Synthese 137*, 273–323.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.

Kolmogorov, A. (1950). *Foundations of the Theory of Probability*. New York: Chelsea Publishing Company. Tr. Nathan Morrison.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, Volume II, pp. 263–93. University of Caifornia Press.

Rényi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Hungarica 6*, 265–333.

Rényi, A. ([1970] 2007a). *Foundations of Probability*. Mineola, NY: Dover Publications, Inc. Reprint of edition published by Holden-Day, 1970.

Rényi, A. ([1970] 2007b). *Probability Theory*. Mineola, NY: Dover publications, Inc. Reprint of edition published by North-Holland, 1970.