Probability All The Way Up

(Or No Probability At All)

David Atkinson and Jeanne Peijnenburg

Faculty of Philosophy, University of Groningen, The Netherlands

21 May 2005

D.Atkinson@rug.nl A.J.M.Peijnenburg@rug.nl

To be published in Synthese

Abstract

Richard Jeffrey's radical probabilism ('probability all the way down') is augmented by the claim that probability cannot be turned into certainty, except by data that logically exclude all alternatives. This claim is illustrated in frequentist language by an infinite nesting of confidence levels, and in Bayesian language by means of Jeffrey's updating of odds.

1 Introduction

Radical probabilism is Richard Jeffrey's emendation of logical positivistic ideas about probability. Rudolf Carnap, the leader of the Vienna Circle and Jeffrey's teacher in postwar Chicago, had maintained that probability judgements consist of two factors: a logical factor involving a prior probability that reflects our ignorance, and an empirical factor composed of experimental evidence. The latter may be described in phenomenalistic or physicalistic language, but the essential feature is its trustworthiness. It is the solid bedrock upon which, using the ignorance prior, our body of knowledge is erected in a process of conditionalization or updating.

His admiration for Carnap notwithstanding, Jeffrey felt unhappy with this view. Firstly he had his doubts about the non-empirical ignorance prior, realising that the choice of a probability distribution always relies on some sort of background knowledge; secondly he felt misgivings about the supposedly trustworthy character of the empirical component, fearing that certitude is not to be attained by mortals. Especially the latter feelings led him to challenge the Carnapian view, which he termed 'dogmatic probabilism'. Dogmatic, because it proclaims that we attach probability values only on the ground of certainties, and hence that we learn only by conditioning on the basis of sensation or observation statements that themselves remain unquestioned.

Radical probabilism is Jeffrey's answer to dogmatic probabilism: it claims that probability values need not be grounded on a certain basis. Its core is the so-called 'probability kinematics', a form of updating based on the familiar Bayesian conditionalization. Suppose we have a hypothesis H, to which we have given a prior probability, P_{old}, and for which new evidence E has just come in. Then according to Bayes, the posterior probability of H, $P_{\mathsf{new}}(H)$, is calculated in the following way:

$$P_{\mathsf{new}}(H) \equiv P(H|E) = P(E|H)P_{\mathsf{old}}(H)/P(E)\ , \tag{1}$$

where $P(E)$ can be written as

$$P(E) = P(E|H)P_{\mathsf{old}}(H) + P(E|\neg H)(1 - P_{\mathsf{old}}(H))\ ,$$

a consequence of the axioms of probability theory known as Jeffrey conditionalization. Instead of working with probabilities directly, Jeffrey often prefers to use odds, i.e. ratios of probabilities of hits (favourable cases) over probabilities of misses (unfavourable cases). He writes

$$\frac{\mathsf{new\,odds}(H)}{\mathsf{old\,odds}(H)} = \frac{P(E|H)}{P(E|\neg H)}\ , \tag{2}$$

showing that the supposedly solid basis, E, enters as a factor, which multiplies the old odds to yield the new ones. Hence, although Eq.(1) and Eq.(2) are mathematically equivalent (as will be shown in Sect. 3), the latter has an important advantage over former. For it frees us from the idea that the updating of probability values can only take place on the basis of information, E, that is taken to be certain. Moreover, Eq.(2) makes it easier to distinguish between other people's observations, which one might adopt, and their prior judgemental states, for which one could substitute one's own (Jeffrey 2001, 6-8).

Jeffrey often contrasted his views with those of C.I. Lewis, whose ideas on probability in this respect resemble Carnap's. During a talk in 2001 in Siena, Jeffrey quoted Lewis as follows:

> If anything is to be probable, then something must be certain. The data which themselves support a genuine probability, must themselves be certainties. We do have such absolute certainties, in the sense data initiating belief and in those passages of experience which may later confirm it.
> (Lewis 1946, 186; quoted in Jeffrey 2001, 11).

Jeffrey's radical probabilism denies precisely this claim. It disavows the idea that probabilitities are grounded in certainties under the motto: "it can be probabilities all the way down, to the roots" (Jeffrey 2001, 1; Jeffrey 1992, 11).

In this article we propose to complement Jeffrey's claim. For in saying that probabilities need not be grounded in certainties, Jeffrey told only half of the story, thus revealing only one side of a balanced probabilistic view of the world. The other half concerns updating on the basis of evidence, where this evidence may be grounded in certainty, or, as Jeffrey stresses, in uncertainty. Once we assign mere probabilities on the basis of (certain or uncertain) evidence, no amount of updating will free us from the treadmill of probabilistic statements: it is impossible to achieve certainty. This impossibility is not an empirical matter, stemming from the hard-gained lessons of experience. Nor is it a pragmatic or merely technical impossibility, arising from the fact that life is short and our equipment inadequate. No, it is *logically* impossible to go from probability to certainty, on pain of forsaking the concept of 'probability' itself. In this sense it is not only 'probability all the way down', seeking evidence for evidence and stopping on a basis that is itself uncertain, but also 'probability all the way up', seeking ever finer probability values and accepting the logical impossibility to do better than that.

Our thesis that assigning probabilities has no logical end can be made in the vocabulary of relative frequencies, but also in that of degrees of belief. In both of these cases, we will have either probability all the way up, or no probability at all. To keep our distance from the ongoing debate between objectivists and subjectivists, we make our point first in a language that is mostly employed by champions of the frequency interpretation of probability, viz. that of statisticians concerning confidence levels (Sect. 2). We look at the familiar situation of the repeated tossing of a fair coin, and observe that, in order to make sense of the definition of confidence levels, it is not enough to consider just one finite run of tosses: we must needs consider indefinitely many nestings of runs. In Sect. 3 we consider the testing of a hypothesis, H, on the basis of evidence E, and we do this first by statistical means, using the formalism developed in Sect. 2; and then we explain how the testing works for the Bayesian. In both cases, it is supposed that the evidence, E, is not inconsistent with the negation of the hypothesis, H (which implies $P(E|\neg H) \neq 0$). For if E *were* inconsistent with $\neg H$, then H would necessarily be true, without any probabilistic considerations, whether of statistical or Bayesian kinship. In Sect. 4 we summarize our findings and spell out some philosophical consequences.

2 Confidence Limits

If one tosses a fair coin 1600 times one expects 800 heads on the average. Moreover, if one repeats the run of 1600 tosses many times, one expects that, in 19 out of 20 runs, the actual number of heads will lie between 760 and 840. On the average, in only 5% of the runs will the number of heads be less than 760 or greater than 840. What, in the previous sentences, do the phrases “on the average” and “one expects that” mean for a latter day frequentist?

A naive explication of the meaning of these phrases is that they find their justification in an infinite repetition of the runs of 1600 tosses. The average number of heads, it might be maintained, would then be precisely 800, and the percentage of runs in which the number of heads, h, satisfies $760 \leq h \leq 840$, would be 95.45% (this percentage corresponds to two standard deviations of the normal distribution). This explication is unsatisfactory, however, for the reference to an infinite repetition is illicit: one can at best appeal to the limit of larger and larger numbers of repetitions. For a finite number of repetitions, the frequentist might like to say it is *very likely* that in 95% of the runs, the number of heads will lie between 760 and 840. But what can it mean for the frequentist to say that something is very likely, or that it has a high probability? How can he employ these terms without circularity?

A way out of the frequentist's quandary is to invoke the concept of the *nesting* of repetitions of repetitions of runs. One defines a hierarchy of sets in the following way: suppose that the run of 1600 tosses is itself repeated 1600 times. Instead of saying that we *expect* the number of heads to lie between 760 and 840 in 95% of the 1600 runs, we rather divide the set of runs into a subset, called good runs, in which $h \in \{760, 840\}$, and its complement, called bad runs, in which $h \notin \{760, 840\}$. Let g_1 be the number of these good runs. How do we give meaning to the vague expectation that $g_1 \approx 1520$, i.e. that about 95% of the 1600 runs are good? After all, we will not inevitably find that 1520 of the 1600 runs are good – the number might turn out to be slightly smaller, or slightly larger. Indeed, while the mean number of good runs is $\mu(g_1) = 1600 \times 0.95 = 1520$, the variance is $\sigma^2(g_1) = 1600 \times 0.95 \times 0.05 = 76$, so the standard deviation itself is about 8.5. Therefore $1503 \leq g_1 \leq 1537$ at the 95% level of confidence. We must in turn imbue this statement with meaning.

We can *repeat* the 1600 repetitions of 1600 throws 1600 times. The elements of the new repetition, let us call it rep[2], are the repetitions, rep[1], of the runs of throws. Let us call a given rep[1] good if the number, g_1, of good runs it contains satisfies $1503 \leq g_1 \leq 1537$, and suppose that the number of these good repetitions (rep[1]) is g_2. Since rep[2] contains 1600 repetitions (rep[1]), the mean and variance of g_2 are $\mu(g_2) = 1600 \times 0.95 = 1520$ and $\sigma^2(g_2) = 1600 \times 0.95 \times 0.05 = 76$, i.e. g_2 has the same mean and variance in rep[2] as g_1 has in rep[1]. So $1503 \leq g_2 \leq 1537$ at the 95% level of confidence, and it is now obvious how this statement is to be given meaning, namely by a further nesting in a new repetition — rep[3] — of rep[2] repetitions. A good rep[2] is defined to be one in which $1503 \leq g_2 \leq 1537$, and if g_3 is the number of them, then $1503 \leq g_3 \leq 1537$ at the 95% level of confidence, which in turn is given meaning in a similar way at the next level of the hierarchy, and so on *ad infinitum.*

We do not claim that an experimenter must actually indulge in a nesting of confidence levels, let alone an infinite one. We have merely tried to formulate, on behalf of the frequentist as it were, a sort of justification for the use of confidence testing itself. Apart from the question as to whether this justification, or indeed the method itself, is thought to be kosher or not, our point is that the appeal to infinite nesting is *inescapable.* For if it is uncertain whether any toss of the coin will result in a head or a tail, then it is logically impossible for the nesting of sets to terminate. To demonstrate this fact, we show first that a contradiction would ensue if the nesting were to terminate at the first level. Having introduced the idea, we then generalize it to the case that termination occurs at any level at all.

Suppose first that it is not merely *likely* that the number of good runs in the first repetition, rep[1], satisfies $1503 \leq g_1 \leq 1537$, but rather it is *certain* that g_1 lies in this interval. Then no further nesting would be required; but it could be, in a given experiment, that the first 1599 repetitions of 1600 tosses have only produced 1502 good runs. Since there must be at least 1503 good runs, according to the supposition, the last repetition must of necessity be good, which means that the number of heads produced in this last run must satisfy $760 \leq g_1 \leq 840$. However, the first 1599 tosses in this last run might have produced only 759 heads, and since the run must be good, the last throw *must result in a head.* But this is inconsistent with the fact that any particular toss could give a tail.

We now generalize the proof to the case that the nesting terminates at *any* finite level, rather than the first one. If there exists an nth-order nesting, at which the number, g_n, of good repetitions (rep$[n-1]$) must satisfy $1503 \leq g_n \leq 1537$ precisely, then a contradiction ensues with the assumption that it is uncertain whether any toss will produce a head. To show this, we suppose *per impossibile* that $1503 \leq g_n \leq 1537$ must hold. It could still happen in an experiment that, among the first 1599 of the repetitions (rep$[n-1]$), only 1502 were good, so the above supposition would then imply that the 1600th repetition (rep$[n-1]$) must be good. But it could also be the case that this repetition itself is such that, of its first 1599 repetitions (rep$[n-2]$), only 1502 are good, so that the 1600th repetition (rep$[n-2]$) must be good, since otherwise the 1600th repetition (rep$[n-1]$) would not be good, whereas in the situation sketched, and according to the supposition, it must be good. Clearly this reasoning can be iterated back to the second repetition (rep[2]), in which it could similarly devolve that the 1600th repetition (rep[1]) must be good, so then the number of heads thrown in this last run must satisfy $760 \leq h \leq 840$. As before, it could be that only 759 heads have been produced in the first 1599 throws of this 1600th run of rep[1], and this means that the last throw must result in a head. But this is inconsistent with $P(h) < 1$. The original supposition, that there exists an nth-order, at which the nesting terminates, has thus produced a contradiction, so it must be discarded.

Hence the infinite nesting is a *sine qua non* of the uncertainty in the results of the elementary trials (the tosses of the coin). If there is probability at the basic level (other than zero or one), then there is probability at all higher levels. In this sense, frequentism implies probability all the way up. We will now show that a similar implication holds in Bayesianism too.

3 Bayesian Updating

Suppose that there are two machines, each producing trick coins that are biased. We imagine that the first machine is so adjusted that it produces coins that are biased in such a way that, on tossing one of them, the probability of a head is $\frac{3}{4}$, that of a tail $\frac{1}{4}$. The second machine produces coins such that the probability of a head is only $\frac{2}{5}$, that of a tail $\frac{3}{5}$. Suppose that all the coins produced by the two machines in a day are mixed in a bag, and that one coin is selected from it. Does it come from the first or the second machine?

The statistician, using confidence limits as in the previous section, would throw the coin many times, let us say 1600 times, and note the number of heads, h, resulting from the 1600 tosses of the coin. If the coin came from the first machine, the mean value for h is 1200, but of course the actual value will in general deviate from this value, so the statistician will also calculate the standard deviation, $\sqrt{1600 \times \frac{3}{4} \times \frac{1}{4}} \approx 17$. At the 95% level of confidence, it is expected that h will lie within two standard deviations of 1200, i.e. it will lie between 1166 and 1234.

If the coin came from the second machine, however, the mean value for h is now only 640, and the standard deviation is $\sqrt{1600 \times \frac{2}{5} \times \frac{3}{5}} \approx 19.5$. At the 95% level of confidence, it is in this case expected that h will lie within twice this standard deviation of 640, i.e. it will lie between 601 and 679.

At a confidence level of 95%, it is easy to decide which machine produced the coin, since the two standard deviation intervals do not overlap. Indeed they do not do so at the 99% confidence level, corresponding to three standard deviations either. By nesting runs, one can increase the probability that the coin came from one machine, rather from the other, but clearly certainty — probability 1 — can never be achieved.

What would a subjectivist, in this case a Bayesian, make of this problem? For him, probabilities are degrees of belief in alternative options or hypotheses. In our example, there are two hypotheses, namely H_1, that the coin came from the first machine, and H_2, that it came from the second machine. The initial degree of belief that the coin came from the first machine is changed, or updated, on the basis of evidence, the result of 1600 tosses of the coin, in accordance with the formula

$$P(H_1|E) = \frac{P(E|H_1)P_0(H_1)}{P(E|H_1)P_0(H_1) + P(E|H_2)P_0(H_2)}\,. \tag{3}$$

Here $P(H_1|E)$ is the conditional probability that hypothesis H_1 is true, given the evidence, E, namely that h heads resulted from a run of 1600 tosses (that is, the conditional probability that the coin came from the first machine). $P(E|H_1)$ is the probability of scoring h heads in 1600 tosses of a coin from the first machine, while $P(E|H_2)$ is the corresponding probability of scoring h heads in 1600 tosses of a coin from the second machine. Finally, $P_0(H_1)$ is the initial degree of belief that the coin came from the first machine, and $P_0(H_2)$ is the initial degree of belief that it came from the second machine. Since it is supposed that the coin did indeed come from one or other of the machines,

$$P_0(H_1) + P_0(H_2) = 1\,. \tag{4}$$

In parallel with the updating (3) of the degree of belief that the coin came from the first machine, we have

$$P(H_2|E) = \frac{P(E|H_2)P_0(H_2)}{P(E|H_1)P_0(H_1) + P(E|H_2)P_0(H_2)}\,, \tag{5}$$

as the updating of the degree of belief that the coin came from the second machine, and of course $P(H_1|E) + P(H_2|E) = 1$. If the Bayesian has no reason to give more credence to one supposition rather than the other, he will make the prior probabilities equal: $P_0(H_1) = P_0(H_2) = \frac{1}{2}$, but this is not necessary. He may have extra information, for example that the first machine was turned on an hour before the second, while they were turned off together, and in this case he may well choose $P_0(H_1)$ to be greater than $\frac{1}{2}$.

Following Jeffrey, we shall for convenience rewrite Eq.(3) in terms of odds rather than probabilities. The odds associated with a hypothesis are defined as the ratio of the probability that the hypothesis is true to the probability that it is false. Thus the prior odds, $\Omega_0(H_1)$, for hypothesis 1 are

$$\Omega_0(H_1) = \frac{P_0(H_1)}{P_0(H_2)} = \frac{P_0(H_1)}{1 - P_0(H_1)}\,. \tag{6}$$

The updated, or posterior odds are similarly

$$\Omega(H_1|E) = \frac{P(H_1|E)}{P(H_2|E)} = \frac{P(H_1|E)}{1 - P(H_1|E)}\,. \tag{7}$$

On dividing Eq.(3) by Eq.(5), and with use of Eq.(6) and Eq.(7), we obtain

$$\Omega(H_1|E) = \frac{P(E|H_1)}{P(E|H_2)}\,\Omega_0(H_1)\,. \tag{8}$$

As Jeffrey has noted, such an expression for posterior odds separates neatly into a factor that is uniquely determined by the evidence (in this case the number of heads in 1600 throws), and the subjective prior odds. Note that while Jeffrey's expression (8) has been derived from the Bayes formula (3), the derivation can be equally well run in the opposite direction. In other words, Jeffrey's updating of odds is precisely equivalent to Bayes' updating of probabilities.

As intimated earlier, we assume that the evidence E is not flatly inconsistent with H_2, for that would logically imply H_1, without any probabilistic considerations. From Eq.(8) we observe that the posterior odds can only be infinite if $\Omega_0(H_1) = \infty$, since $P(E|H_2) \neq 0$ by assumption. For the case of the coin and the two machines in particular, $P(E|H_2)$ is nonzero for any h between 0 and 1600 inclusive. So the only way that one can achieve an infinite posterior odds is to start with infinite prior odds. In other words, the posterior probability can only be one if the prior probability is one. Just as it is logically impossible to achieve certainty by the nesting of confidence levels when one starts from an uncertain hypothesis, so it is logically impossible by Bayesian updating to achieve certainty from an uncertain prior.

4 Envoi

We have seen that Jeffrey's 'probability all the way down' may be augmented by another adage, 'probability all the way up'. The two maxims are by no means the same. Whereas the former states that the base of our knowledge can be uncertain to its deepest roots, the latter implies that once we start assigning probability values to hypotheses, we will never achieve certainty either. Jeffrey's motto can be seen as an argument against foundationalism in epistemology, at least in the form that occurs in for example C.I. Lewis's work. Our adage 'probability all the way up', on the other hand, functions more as an argument for the thesis that the difference between 'probable' and 'certain' (or between inductive and deductive reasoning) is strict rather than gradual. No matter how many times, on the basis of incoming information, we refine probability values, it is logically impossible to achieve certainty. For if in due course we *were* to reach certainty, then the whole enterprise of giving probability values to hypotheses would collapse at the lowest level, as we have shown. This is the reason for our claim 'probability all the way up or no probability at all', and we observe here another distinction between Jeffrey's view and ours. As Jeffrey admits, it is perfectly *possible* to update probability values from a basis that is taken to be certain, i.e. it *can* be probability all the way down, but it *need* not be so. Probability all the way up, on the other hand, *must* be endorsed, on pain of deserting the field of probability altogether.

Despite their differences, the two views of probability, all the way down and all the way up, round out the account. Taken together, they complete the story and substantiate the balanced probabilistic world-view that Hans Reichenbach surely had in mind when he wrote "all we have is an elastic net of probability connections, floating in open space." (Reichenbach 1938, 192).

Acknowledgements
We should like to thank Barteld Kooi, Theo Kuipers and Jan Willem Romeijn for useful discussions stimulated by an earlier version of this paper, and an anonymous referee for his or her critical and constructive comments.

References

Jeffrey, R.C., 1992, *Probability and the Art of Judgement.* Cambridge: Cambridge U.P.

Jeffrey, R.C., 2001, 'Epistemology Probabilized', in: Johan van Benthem (ed.), *Theoretical Aspects of Rationality and Knowledge.* Proceedings of the Eighth Conference TARK 2001, July 8-10, 2001, Siena. San Francisco: Morgan Kaufmann Publishers.

Lewis, C.I., 1946, *An Analysis of Knowledge and Valuation.* Illinois: Open Court.

Reichenbach, H., 1938, *Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge.* Chicago: University of Chicago Press. Seventh impression, 1970.