PART ONE

Probability

1

Symmetry and Its Discontents

The following paper consists of two parts. In the first it is argued that Bruno de Finetti's theory of subjective probability provides a partial resolution of Hume's problem of induction, if that problem is cast in a certain way. De Finetti's solution depends in a crucial way, however, on a symmetry assumption – exchangeability – and in the second half of the paper the broader question of the use of symmetry arguments in probability is analyzed. The problems and difficulties that can arise are explicated through historical examples which illustrate how symmetry arguments have played an important role in probability theory throughout its development. In a concluding section the proper role of such arguments is discussed.

1. THE DE FINETTI REPRESENTATION THEOREM

Let $X_1, X_2, X_3, \ldots$ be an infinite sequence of 0,1-valued random variables, which may be thought of as recording when an event occurs in a sequence of repeated trials (e.g., tossing a coin, with 1 if heads, 0 if tails). The sequence is said to be *exchangeable* if all finite sequences of the same length with the same number of ones have the same probability, i.e., if for all positive integers n and permutations σ of $\{1, 2, 3, \ldots, n\}$,

$$\begin{aligned} P[X_1 = e_1, X_2 = e_2, \ldots, X_n = e_n] \\ = P[X_1 = e_{\sigma(1)}, X_2 = e_{\sigma(2)}, \ldots, X_n = e_{\sigma(n)}], \end{aligned}$$

where e_i denotes either a 0 or a 1. For example, when $n = 3$, this means that

$$\begin{aligned} P[1, 0, 0] = P[0, 1, 0] = P[0, 0, 1] \quad \text{and} \\ P[1, 1, 0] = P[1, 0, 1] = P[0, 1, 1]. \end{aligned}$$

(Note, however, that $P[1, 0, 0]$ is not assumed to equal $P[1, 1, 0]$; in general, these probabilities may be quite different.)

Reprinted with permission from Brian Skyrms and William L. Harper (eds.), *Causation, Chance, and Credence* 1 (1988): 155–190, © 1988 by Kluwer Academic Publishers.

In 1931 the Italian probabilist Bruno de Finetti proved his famous *de Finetti Representation Theorem. Let* $X_1, X_2, X_3, \ldots$ *be an infinite exchangeable sequence of* 0,1*-valued random variables, and let* $S_n = X_1 + X_2 + \cdots + X_n$ *denote the number of ones in a sequence of length n. Then it follows that:*

1. *the limiting frequency* $Z =: \lim_{n\to\infty}(S_n/n)$ *exists with probability 1.*
2. *if* $\mu(A) =: P[Z \in A]$ *is the probability distribution of Z, then*

$$P[S_n = k] = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \mathrm{d}\mu(p)$$

for all n and k.[1]

This remarkable result has several important implications. First, contrary to popular belief, subjectivists clearly believe in the existence of infinite limiting relative frequencies – at least to the extent that they are willing to talk about an (admittedly hypothetical) infinite sequence of trials.[2] The existence of such limiting frequencies follows as a purely mathematical consequence of the assumption of exchangeability.[3] When an extreme subjectivist such as de Finetti denies the existence of objective chance or physical probability, what is really being disputed is whether limiting frequencies are objective or physical properties.

There are several grounds for such a position, but all center around the question of what "object" an objective probability is a property of. Surely not the infinite sequence, for that is merely a convenient fiction (Jeffrey 1977). Currently the most fashionable stance seems to be that objective probabilities are a *dispositional property* or *propensity* which manifests itself in, and may be measured with ever-increasing accuracy by, finite sequences of ever-increasing length (e.g., Kyburg 1974).

But again, a property of what? Not the coin, inasmuch as some people can toss a so-called fair coin so that it lands heads 60% of the time or even more (provided the coin lands on a soft surface such as sand rather than a hard surface where it can bounce). Some philosophers attempt to evade this type of difficulty by ascribing propensities to a *chance set-up* (e.g., Hacking 1965): in the case of coin-tossing, the coin *and* the manner in which it is tossed. But if the coin were indeed tossed in an identical manner on every trial, it would always come up heads or always come up tails; it is precisely because the manner in which the coin is tossed on each trial is *not* identical that the coin can come up both ways. The suggested chance set-up is in fact nothing other than a sequence of objectively differing trials which we are subjectively

unable to distinguish between. At best, the infinite limiting frequency is a property of an "object" enjoying both objective and subjective features.

2. DE FINETTI VANQUISHES HUME

The most important philosophical consequence of the de Finetti representation theorem is that it leads to a solution to *Hume's problem of induction:* why should one expect the future to resemble the past? In the coin-tossing situation, this reduces to: in a long sequence of tosses, if a coin comes up heads with a certain frequency, why are we justified in believing that in future tosses of the same coin, it will again come up heads (approximately) the same fraction of the time?

De Finetti's answer to this question is remarkably simple. Given the information that in n tosses a coin came up heads k times, such data is incorporated into one's probability function via

Bayes's rule of conditioning: $P[A\,|B] = P[A \text{ and } B]/P[B]$.

If n is large and $p^* = k/n$, then – except for certain highly opinionated, eccentric, or downright kinky "priors" $\mathrm{d}\mu$ – it is easy to prove that the resulting posterior probability distribution on p will be highly peaked about p^*; that is, the resulting probability distribution for the sequence of coin tosses looks approximately like (in a sense that can be made mathematically precise) a sequence of independent and identically distributed Bernoulli trials with parameter p^* (i.e., independent tosses of a p^* coin). By the weak law of large numbers it follows that, with high probability, subsequent tosses of the coin will result in a relative frequency of heads very close to p^*.

Let us critically examine this argument. Mathematically it is, of course, unassailable. It implicitly contains, however, several key suppositions:

1. P is operationally defined in terms of betting odds.
2. P satisfies the axioms of mathematical probability.
3. P is modified upon the receipt of new information by Bayesian conditioning.
4. P is assumed to be exchangeable.

In de Finetti's system, degree of belief is quantified by the betting odds one assigns to an event. By a Dutch book or coherence argument, one deduces that these betting odds should be consistent with the axioms of mathematical probability. Conditional probabilities are initially defined in terms of conditional bets and Bayes's rule of conditioning is deduced as a consequence of coherence. The relevance of conditional probabilities to inductive inference

is the *dynamic assumption of Bayesianism* (Hacking 1967): if one learns that B has occurred, then one's new probability assignment is $P[A\,|B]$. In general, however, conditional probabilities can behave in very nonHumeian ways, and (infinite) exchangeability is taken as describing the special class of situations in which Humeian induction is appropriate.

This paper will largely concern itself with the validity of this last assumption. Suffice it to say that, like Ramsey (1926), one may view the subjectivist interpretation as simply capturing *one* of the many possible meanings or useages of probability; that the Dutch book and other derivations of the axioms may be regarded as plausibility arguments (rather than normatively compelling); and that although a substantial literature has emerged in recent decades concerning the limitations of Bayesian conditioning, the difficulties discussed and limitations raised in that literature do not seem particularly applicable to most of the situations typically envisaged in discussions of Hume's problem.

The assumption of exchangeability, however, seems more immediately vulnerable. Isn't it essentially circular, in effect assuming what one wishes to prove? Of course, in one sense this must obviously be the case. All mathematics is essentially tautologous, and any implication is contained in its premises. Nevertheless, mathematics has its uses. Formal logic and subjective probability are both theories of consistency, enabling us to translate certain assumptions into others more readily palatable.

What de Finetti's argument really comes down to is this: if future outcomes are viewed as exchangeable, i.e., no one pattern is viewed as any more or less likely than any other (with the same number of successes), then when an event occurs with a certain frequency in an initial segment of the future, we must, if we are to be consistent, think it likely that that event will occur with approximately the same frequency in later trials. Conversely, if we do not accept this, it means that we must have – prospectively – thought certain patterns more likely than others. Which means that we must have possessed more information than is ordinarily posited in discussions of Humeian induction.

And there the matter would appear to stand. Or does it?

3. THE INSIDIOUS ASSUMPTION OF SYMMETRY

Exchangeability is one of many instances of the use of symmetry arguments to be found throughout the historical development of mathematical probability and inductive logic. But while such arguments often have a seductive attraction, they also often carry with them "hidden baggage": implications

or consequences, sometimes far from obvious, which later cast serious doubt on their validity. We will discuss three historically important examples, all involving attempts to justify induction by the use of probability theory, and all (in effect) involving the appropriate choice of prior $\mathrm{d}\mu$ in the de Finetti representation.

Example 3.1. *Bayes's argument for the Bayes–Laplace prior.*

Consider "an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it" (Bayes 1764). Implicitly invoking a symmetry argument, Bayes argued that "concerning such an event I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another," i.e., that in a sequence of n trials one's probability assignment for S_n, the number of heads, should satisfy

Bayes's Postulate: $P[S_n = k] = 1/(n+1)$.

That is, the number of heads can assume any of the $n+1$ values $0, 1, 2, \ldots, n$ and, absent further information, all $n+1$ values are viewed as equally likely. In a famous Scholium, Bayes concluded that if this were indeed the case, then the prior probability $\mathrm{d}\mu(p)$ must be the "flat" prior $\mathrm{d}p$.[4]

Although Bayes's exact reasoning at this point is somewhat unclear, it can easily be made rigorous: Taking $k = n$ in the de Finetti representation and using Bayes's postulate, it follows that

$$\int_0^1 p^n \mathrm{d}\mu(p) = 1/(n+1).$$

The integral on the left-hand side is the n-th moment of $\mathrm{d}\mu$, so Bayes's assumption uniquely determines the moments of $\mathrm{d}\mu$. But since $\mathrm{d}\mu$ is concentrated on a compact set, it follows by a theorem of Hausdorff that $\mathrm{d}\mu$, if it exists, is in turn determined by its moments. That is, there can be at most one probability measure $\mathrm{d}\mu$ which satisfies Bayes's assumption $P[S_n = k] = 1/(n+1)$. But the flat measure $\mathrm{d}p$ does satisfy this integral equation, i.e.,

$$\int_0^1 p^n \mathrm{d}p = 1/(n+1),$$

hence $\mathrm{d}\mu$ must be $\mathrm{d}p$.

Bayes's argument is quite attractive. A modern-day subjectivist might view Bayes's assumption as a *definition* (possibly one of many) of "complete

ignorance" (rather than consider "complete ignorance" to be an *a priori* meaningful concept), but would probably find Bayes's argument otherwise unobjectionable.

The argument in its original form, however, did not go uncriticized. As Boole (1854, pp. 369–375) noted, rather than consider the events $[S_n = k]$ to be equally likely, one could equally plausibly take all sequences of a fixed length (or "constitutions") to be so. Thus, for $n = 3$

$$P[000] = P[100] = P[010] = P[001] = P[110] = P[101] = P[011] = P[111] = 1/8.$$

To many, this assignment seemed a far more natural way of quantifying ignorance than Bayes's.

Unfortunately, it contains a time-bomb with a very short fuse. As Carnap (1950, p. 565) later noted (and Boole himself had already remarked), this probability assignment corresponds to *independent* trials, and thus remains unchanged when conditioned on the past, an obviously unsatisfactory choice for modeling inductive inference, inasmuch as "past experience does not in this case affect future expectation" (Boole 1854, p. 372).

In his *Logical Foundations of Probability* (1950), Carnap announced that in a later volume, "a quantitative system of inductive logic" would be constructed, based upon a function Carnap denoted c^*. Carnap's c^* function was, in effect, the one already proposed by Bayes. But Carnap grew uneasy with this unique choice, and in his monograph *The Continuum of Inductive Methods* (1952), he advocated instead the use of a one-parameter family containing c^*. Unknown to Carnap, however, he had been anticipated in this, almost a quarter of a century earlier, by the English philosopher William Ernest Johnson.

Example 3.2. *W. E. Johnson's sufficientness postulate.*

In 1924 Johnson, a Cambridge logician, proposed a multinomial generalization of Bayes's postulate. Suppose there are $t \geq 2$ categories or types, and in n trials there are n_1 outcomes of the first type, n_2 outcomes of the second type, ..., and n_t outcomes of the t-th type, so that $n = n_1 + n_2 + \cdots + n_t$. The sequence $(n_1, n_2, \ldots, n_t)$ is termed an *ordered t-partition of n*. Bayes had considered the case $t = 2$, and his postulate is equivalent to assuming that all ordered 2-partitions $(k, n - k)$ are equally likely. Now Johnson proposed as its generalization

Johnson's combination postulate: Every ordered t-partition of n is equally likely.

For example, if $t = 3$ and $n = 4$, then there are 15 possible ordered 3-partitions of 4, viz.:

n_1	n_2	n_3
4	0	0
3	1	0
3	0	1
2	2	0
2	1	1
2	0	2
1	3	0
1	2	1
1	1	2
1	0	3
0	4	0
0	3	1
0	2	2
0	1	3
0	0	4

and each of these is assumed to be equally likely.

Johnson did not work with integral representations but, like Carnap, with finite sequences. In so doing he introduced a second postulate, his "permutation postulate." This was none other than the assumption of exchangeability, thus anticipating de Finetti (1931) by almost a decade! (If one labels the types or categories with the letters of a t-letter alphabet, exchangeability here means that all words of the same length, containing the same number of letters of each type, are equally likely). Together, the combination and permutation postulates uniquely determine the probability of any specific finite sequence. For example, if one considers the fifth partition in the table above, $4 = 2 + 1 + 1$, then there are twelve sequences which give rise to such a partition, viz.

x_1	x_2	x_3	x_4
1	1	2	3
1	1	3	2
1	2	1	3
1	2	3	1
1	3	1	2
1	3	2	1
2	1	1	3
2	1	3	1
2	3	1	1
3	1	1	2
3	1	2	1
3	2	1	1

and each of these are thus assumed to have probability $(1/15)(1/12) = 1/180$. The resulting probability assignment on finite sequences is identical with Carnap's c^*.

Despite its mathematical elegance, Johnson's "combination postulate" is obviously arbitrary, and Johnson was later led to substitute for it another, more plausible one, his "sufficientness postulate." This new postulate assumes for all n

Johnson's sufficientness postulate:

$$P[X_{n+1} = j | X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n] = f(n_j, n).$$

That is, the conditional probability that the next outcome is of type j depends only on the number of previous trials and the number of previous outcomes of type j, but not on the frequencies of the other types or the specific trials on which they occurred. If, for example $t = 3$, $n = 10$, and $n_1 = 4$, the postulate asserts that on trial 11 the (conditional) probability of obtaining a 1 is the same for all sequences containing four 1's and 6 not –1's, and that this conditional probability does not depend on whether there were six 2's and no 3's, or five 2's and one 3, and so on. (Note that the postulate implicitly assumes that all finite sequences have positive probability, so that the conditional probabilities are well-defined.)

Johnson's sufficientness postulate makes what seems a minimal assumption: absence of knowledge about different types is interpreted to mean that information about the frequency of one type conveys no information about the likelihood of other types occurring. It is therefore rather surprising that it follows from the postulate that the probability function P is uniquely determined up to a constant:

Theorem (Johnson 1932). *If P satisfies the sufficientness postulate and $t \geq 3$, then either the outcomes are independent or there exists a $k > 0$ such that*

$$f(n_i, n) = \{n_i + k\}/\{n + tk\}.$$

This is, of course, nothing other than Carnap's "continuum of inductive methods."[5]

The de Finetti representation theorem can be generalized to a much wider class of infinite sequences of random variables than those taking on just two values (e.g., Hewitt and Savage 1955). In the multinomial case now being discussed, the de Finetti representation states that every exchangeable probability can be written as a mixture of multinomial probabilities. Just as Bayes's postulate implied that the prior $d\mu$ in the de Finetti representation was the flat prior, Johnson's theorem implies that the mixing measure $d\mu$ in

the de Finetti representation is the *symmetric Dirichlet prior*

$$\Gamma(tk)/\Gamma(k)^t\, p_1^{k-1} p_2^{k-1} \ldots p_1^{k-1} \mathrm{d}p_1 \mathrm{d}p_2 \ldots \mathrm{d}p_{t-1};$$

a truly remarkable result, providing a subjectivistic justification for the use of the mathematically attractive Dirichlet prior.[6]

Despite its surface plausibility, Johnson's sufficientness postulate is often too strong an assumption. While engaged in cryptanalytic work for the British government at Bletchley Park during World War II, the English logician Alan Turing realized that even if one lacks specific knowledge about individual category types, the frequencies $n_1, n_2, \ldots, n_t$ may contain relevant information about predictive probabilities, namely the information contained in the *frequencies of the frequencies*.

Let a_r = the number of frequencies n_i equal to r; a_r is called the frequency of the frequency r. For example, if $t = 4$, $n = 10$, and one observes the sequence 4241121442, then $n_1 = 3$, $n_2 = 3$, $n_3 = 0$, $n_4 = 4$ and $a_0 = 1$, $a_1 = 0$, $a_2 = 0$, $a_3 = 2$, $a_4 = 1$. (A convenient shorthand for this is $0^1 1^0 2^0 3^2 4^1$.) Although it is far from obvious, the a_r may be used to estimate cell probabilities: see Good (1965, p. 68).[7]

Example 3.3. *Exchangeability and partial exchangeability.*

Given the failure of such attempts, de Finetti's program must be seen as a further retreat from the program of attempting to provide a unique, quantitative account of induction. Just as Johnson's sufficientness postulate broadened the class of inductive probabilities from that generated by the Bayes–Laplace prior to the continuum generated by the symmetric Dirichlet priors, so de Finetti extended the class of possible inductive probabilities even further to include *any* exchangeable probability assignment.

But what of the symmetry assumption of exchangeability? Even this is not immune to criticism (as de Finetti himself recognized). Consider the following sequence: 000101001010100010101001.... Scrutiny of the sequence reveals the interesting feature that although every 0 is followed by a 0 or 1, every 1 is invariably followed by a 0. If this feature were observed to persist over a long segment of the sequence (or simply that 1's were followed by 0's with high frequency), then this would seem relevant information that should be taken into account when calculating conditional, predictive probabilities. Unfortunately, exchangeable probabilities are useless for such purposes: if P is exchangeable, then the conditional probabilities

$$P[X_{n+1} = j | X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n]$$