THE BAYES MIND: FROM THE ST PETERSBURG PARADOX TO THE NEW YORK STOCK EXCHANGE

MIKHAIL MASOKIN

Submitted for a PhD degree in Philosophy

London School of Economics and Political Science

London, 2001

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

The thesis is an exposition and defence of Bayesianism as the preferred methodology of reasoning under uncertainty in social contexts. Chapter 1 gives a general outline of the foundations of probabilistic reasoning, as well as a critical exposition of the main non-Bayesian approaches to probability. After a brief discussion of the formal theory of probability, the thesis examines some non-Bayesian interpretations of the probability calculus, and purports to show their insufficiency.

Chapter 2 provides an outline of the Bayesian (subjectivist) research programme. The opening sections of the chapter contain a historical overview of Bayesianism, as well as a defence of the assumptions on which it rests. The concluding sections then examine some of the key issues of contention between Bayesians and their critics, such as the nature of empirical confirmation and learning from experience.

Since it is the author's contention that any sound methodology should be applicable, if necessary with modifications, across a wide range of contexts, the concluding two chapters make a Bayesian case first in a theoretical, and next in a practical setting. In particular, Chapter 3 discusses the issue of simplicity as a theoretical virtue. It argues that a Bayesian can coherently and successfully account for the structural or formal simplicity of the hypotheses that he entertains, by using his assignments of subjective prior probabilities in the process known as 'Bayesian Conditionalisation'. It also argues that some of the recent criticisms voiced against the Bayesian account of simplicity are inconsistent and/or question-begging. The process known in statistics as 'curve-fitting' provides the material for the discussion.

Finally, Chapter 4 presents an extension of the Bayesian methodology to practical decision-making, using the context of investing activity. It purports to show that the most convincing picture of economic agents' investing behaviour is best explained by assuming that in the course of such behaviour, the agents maximise their expected utility, as is stipulated by the Bayesian decision theory. The argument revolves around the 'Efficient Markets Hypothesis' in the theory of finance, and the conclusion hinges both on the empirical adequacy of the various versions of this hypothesis and on its behavioural underpinnings.

The thesis contains two appendices, intended to illustrate certain points made in the main body. The first appendix is a critical appraisal of a popular non-Bayesian account of causal inference in statistical contexts, with a bearing on the discussion in Chapter 3, while the second appendix provides a real-life illustration of some of the issues raised in Chapter 4. The overall structure of the thesis is intended to show how, from highly plausible assumptions, one can derive a powerful theory of reasoning under uncertainty that faithfully and uniformly represents both the theoretical and the practical concerns of the human mind.

Contents

CONTENTS

CHAPTER 1: BAYESIANISM IN CONTEXT 7

Rationality 7

Probability Calculus 8

Fields 8

Probability function 9

Conditional probability 11

Beyond the axioms 12

Classical Interpretation 15

Independence 15

Counting favourable outcomes 17

Symmetry and ignorance 19

Physicalist Interpretation 22

Von Mises: empiricism 22

Popper: conjectures and refutations 32

Logical Interpretation 43

Keynes: extending logic to uncertainty 43

The principle of indifference revisited 45

Carnap: uncertainty without indifference 46

CHAPTER 2: SUBJECTIVE BAYESIANISM 52

Ramsey: Probability and truth 53

Putting your money where your mouth is 53

The St Petersburg Paradox 54

Extracting "subjective value" 58

de Finetti: From belief to probability 59

Taking a gamble 59

From belief to probability: are we there yet? 67

Contents

Savage **71**
Preference 71
Probability 78
Utility 81

Kinematics of belief **83**

Evidential support **93**
Probabilification 94
Synthetic universal statements 103
Conclusion 114

CHAPTER 3: SIMPLICITY **115**

The empirical advantages of a simplicity criterion **116**
Goodness of fit 116
The 'true' curve 118
Distance from the truth 124

The bearing on Bayesianism **130**
Truth vs. probability 131
A trouble with likelihood? 143
The Principle of Simplicity 146

Beyond Akaike **154**
How simple is simplicity? 154
Cognitive utility 156

Contents

CHAPTER 4: MARKET BEHAVIOUR 159

"Fair price" 159

Martingales 162

Which efficiency? 166

Time is of the essence 171

APPENDICES 174

1. **Cartwright on "statistical discovery"** 174
   Introduction 174
   The "statistical discovery" thesis 176
   The "perfect reliability of measurement" thesis 186

2. **The collapse of LTCM** 187
   Introduction 187
   Hedge funds 187
   The 'Asian flu' 188
   Collapse and rescue 189
   Nobel Prize-winning methodology 192
   The bearing on EMH 195

BIBLIOGRAPHY 197

Contents

LIST OF TABLES

Table 1: Relative frequencies of heads and tails, and their changes on the previous toss 25

Table 2: Probability of a given relative frequency vs. the number of tosses 27

Table 3: The subjective value of a St Petersburg gamble 56

LIST OF FIGURES

Figure 1: Relative frequency of heads and its change on the previous toss 26

Figure 2: Probability of a given relative frequency vs. the number of tosses 28

Figure 3: The subjective value of a St Petersburg gamble 56

Figure 4: Linear vs. cubic approximation of sample data 127

CHAPTER 1: BAYESIANISM IN CONTEXT

RATIONALITY

The central tenet of Bayesianism is that in conditions of uncertainty, a rational individual reasons (or must reason, in order to be rational) by assigning a probability to his various beliefs (or, more accurately, to the ideas that he entertains), and subsequently alters his assignments, as new information becomes available, according to the procedure known as Bayesian Conditionalisation. When expanded from rational thought to rational action, Bayesianism also claims that the individual acts (or must act, in order to be rational), so as to maximize his expected utility. This, as far as it goes, is a fairly complete summary of Bayesianism. However, a number of notions used in this summary require an elucidation.

A rational agent is one that makes the best use of the resources at his disposal in pursuit of his objectives. This is, I hope, uncontroversial, albeit possibly insufficiently informative. This abstract definition of rationality can be expanded along a number of lines, all of which rest on equating irrationality with a waste of the limited stock of intellectual and/or material resources available to the agent:

- **Overt logical inconsistency** is irrational, being analogous to trying to make progress along a road while attempting to turn left and right at the same time[1];

- Unnecessarily giving up a **certain advantage** or agreeing to a **certain disadvantage** is irrational.[2]

[1] It is irrational to *believe* an inconsistency; it may be quite rational to advocate it for *someone else's* consumption.

[2] This view does not preclude phenomena such as altruism or accepting a handicap when playing against a weaker partner, for such putative counterexamples are ruled out by the "unnecessarily" qualification.

Also, the intuition as to what constitutes an efficient and what a wasteful use of resources, is quite firmly ingrained in people's minds. After all, we all have first hand experience in the great battle of survival, and although in humans natural selection has been largely subsumed under social selection, ours is still a world where being exceedingly inefficient greatly compromises one's chances of success. As a result, people are pretty adept at recognising efficiency and wastefulness, at least in their extreme manifestations. (Of course, terms like 'exceedingly' and 'extreme' are inevitably vague, but they usually allow for fairly uncontroversial interpretations.) Hence, a casebook of paradigmatically rational and irrational individuals and patterns of behaviour is readily available to a student of the concept of rationality.

PROBABILITY CALCULUS

Fields

The next notion requiring an elucidation is that of probability. The most rigorous and formal definition is given in Measure Theory (cf. Kolmogorov 1950 and numerous subsequent expositions, e.g. Billingsley 1979). Kolmogorov used the extremely general language of *sets*: although his measure-theoretic approach includes a certain 'preferred interpretation', in principle, various kinds of entities can be represented by sets.

E.g., where a repeatable random experiment is to be modelled, a particular (non-empty) set U can be chosen as the "universal set" whose elements w (or "sample points") are all the possible outcomes of the experiment. Suppose the experiment consists in tossing a coin n times and observing the outcomes. Then U is the 2^n-strong set of n-long sequences consisting of 'heads' and 'tails'. Let n=3. Then U={(hhh), (hht), (hth), (thh), **(htt)**, **(tht)**, **(tth)**, (ttt)} (the sequences containing exactly two tails are highlighted in **bold**). Next, a field F is defined on U. A *field* is any collection of subsets of U that includes U itself and is closed under the operations of intersection, union and complementation with respect to U.[3]

[3] Closure under intersections and under unions are not independent requirements since they can be deduced from each other by De Morgan's Law (given closure under complementation). Closure under complementation also ensures that F includes the empty set.

The elements of F are usually referred to as "events". In the coin-tossing case, there is a unique

$$F=\{\emptyset;\ \{(hhh)\};\ \ldots;\ \{(hhh),(hht)\};\ \ldots;\ \{(hhh),(hht),(hth)\};\ \ldots;\ U\}$$

that includes all the subsets of U, from cardinality 0 to cardinality 8

If the *w*s are viewed as the "atomic" states of affairs, then an event x is the set of all those states of affairs w that are consistent with it. The event x of the coin landing heads up 1/3 of the time is the union of all the n-long sequences where the proportion of heads out of n tosses equals 1/3. Say, n=3. Then U={(hhh), (hht), (hth), (thh), **(htt)**, **(tht)**, **(tth)**, (ttt)}, and F={**(htt)**, **(tht)**, **(tth)**}. Then x is consistent with all and only those outcomes that are typed in bold (there are three of them). Similarly, given U, we can identify every event with a set of *w*s. The events are nothing other than the elements of F. Therefore, in applying probability theory we may refer to sets and events (and further, to the propositions or sentences that describe those events) interchangeably.

Probability function

The elements of F constitute the argument domain of a *probability function.* Probability is any function p defined on F that satisfies the following conditions:

(1) For all x in F, $p(x)\geq 0$ (non-negativity);

(2) $p(U)=1$;

(3) If $x\cap y=0$ (i.e., if the two sets are disjoint), then $p(x\cup y)=p(x)+p(y)$ (additivity);

It is a consequence of (1)-(3) that for *any* two sets,

(3') $p(x\cup y)=p(x)+p(y)-p(x\&y)$;

In addition to (1)-(3), the following equality is normally postulated as the definition of conditional probability:

(4) For all x and y in F, $p(x|y)=p(x\&y)/p(y)$, where $p(y)>0$.

Finally, a probability space is any ordered triple $\{U, F, p\}$ that satisfies the above conditions (1)-(4). Some consequences of the axioms of probability are given below.

- No event can be less probable than 0 (or, as follows from axioms (1) and either (2) or (3), more probable than 1). Any statement that includes reference to negative probability or to being "120% sure" must be viewed as a poetic exaggeration.

- The probabilities of any event and its contradictory must add up to 1. *Example*: the probability of a natural number being either odd or even is one. If the sum of two probabilities adds up to other than unity, the corresponding events cannot be genuine contradictories, such as "being tall" and "being short" (one may also be of medium height).

The arguments of a probability function (elements of F) may be referred to, in technical literature, as 'events', but do they fit the more conventional understanding of 'event'? It is easy to show that 'ordinary events' do lend themselves to set-theoretic language, and therefore expressions like 'intersection of two events' are meaningful. Indeed, logicians have long been depicting set-theoretic operations on events, in the shape of Venn diagrams. Speaking of 'ordinary events', we concentrate on what happens within a specific slice of the spatio-temporal continuum.[4] All the different ways in which the continuum is arranged *outside* the relevant 'slice', are subsumed under the same 'event', as elements of the same set. Let us take the 'space' of the football results table in the English Premier League. Since the Premiership includes 20 teams, the event of one of them, say Arsenal FC, winning the League is the set whose elements are the permutations of the other nineteen clubs in places

[4] At a pinch, this idea can be extended to non-physical spaces, such as the space of the football Premiership results table.

from 2 to 20. From now on, I shall feel free to apply set-theoretic language to 'ordinary' events without further justification.

Clearly, since *any* pair of disjoint events exhibits additivity, by mathematical induction this property can be proven for a union of any finite length ("finite additivity"). If, further, field F is closed under infinitely countable unions (i.e., is a sigma-field), one can define probability functions on that field that posses the property of countable additivity. That is, additivity on a sigma-field may hold for *countably infinite* families of events.[5] On purely formal grounds, there is no compelling reason to decide between the weaker (finite) and the stronger (countable) versions of additivity, but certain *interpretations* of the probability calculus may support one but not the other.

Conditional probability

According to (4), the probability of one event *conditional* on another equals the probability of the intersection of the two events divided by the probability of the second event.

Informally, conditional probability is the probability of an event "provided"/"given that"/"in the case where" the other event occurs with certainty (perhaps counterfactually). *Examples*: say, we are interested in the probability of Arsenal winning the Premiership provided it wins its next game. If the probability of its winning *both* the next game and the Premiership is 0.56, and of winning that game - 0.8, then the conditional probability of winning the Premiership is 0.56/0.8=0.7. If, further, the *unconditional* probability of Arsenal winning the league is 0.5 (which is different from its conditional probability), then the event of winning the Premiership is *not* "probabilistically", or "statistically", independent of the event of winning the next game. (Two events are probabilistically independent **iff** the probability of one conditional on the other equals its unconditional probability.) On the other hand, if Arsenal wins the Premiership conditionally on Watford

[5] Closure with respect to countable unions is a necessary but not sufficient condition for a probability function defined on a field exhibiting countable additivity. One has to postulate that property separately.

being promoted from the First League to the Premiership League for the next season with probability 0.5, which equals its unconditional probability, then the two events are probabilistically independent. Another expression of probabilistic independence is that the probability of a conjunction equals the product of probabilities. Indeed,

$p(x) = p(x|y)$ By definition of independence

$p(x|y) = p(x\&y)/p(y)$, where $p(y)>0$ By definition of conditional probability

Therefore,

$p(x\&y) = p(x)*p(y)$.

It should be noted that on competing interpretations, conditional probability is either *defined* by (4), or merely *measured* by it. In the former case, (4) is analytic, in the latter - synthetic. This difference may be important in the context of belief revision.

Which option to take depends on the particular interpretation of probability that seems the most relevant in a given line of inquiry. Either way, the notion of probabilistic independence is crucial as said independence is often taken to be evidence for an absence of causal links between two events, which is why measuring all sorts of probabilities is an important tool in exploring the causal structures of various subject matters.[6]

Beyond the axioms

Lebesgue measure

Some of the interesting subject matters are partitioned in a way that renders the conceptual apparatus of probability axioms insufficient. Many of the applications of probability theory, particularly in physics, require that the probability space be represented as a generalisation of the 3-dimensional Euclidean space $\mathbf{R}^3$ to n dimensions (where $\mathbf{R}$ is the set

[6] However, one ought to be cautious about the extent to which measuring probabilities can result in a *discovery* of causal structures. For further discussion, see Appendix 2.

of real numbers). Such probability spaces $\mathbf{R}^n$ are often referred to as 'geometric probabilities', and usually the most convenient way of imposing a metric (i.e., a probability function) on them is by using a *Lebesgue measure*. A Lebesgue measure is a generalisation of the notions of length (for a line), area (for a surface) and volume (for Euclidean space) to n dimensions. Unfortunately, a Lebesgue measure assigns every point (and therefore any countable set of points), as well as any n-1 dimensional subspace, of an n-dimensional space probability zero.[7]

For instance, let a real number between 0 and 10 be randomly[8] chosen. Real numbers are uncountable (i.e., they cannot be put in a one-to-one correspondence with the set of natural numbers). Therefore using a Lebesgue measure implies that the probability of choosing one of them (say, e - the base of the natural logarithm) is zero. Of course, *one* of these numbers will *definitely* be chosen (i.e., with probability one), but this cannot be achieved by summing up all the (zero) probabilities corresponding to the individual numbers. However, there must be *something* to which non-zero probability could be assigned in a way that, by the property of additivity, guarantees selection from the interval (0, 10). If we partition this "universal" interval into sub-intervals of non-zero length, then all their unions will possess non-zero probability. Therefore, one can measure the probability of randomly selecting a real number situated *between* two other real numbers, for that probability is nothing other than the probability of selecting the interval bounded by those two numbers. Say, if we partition the interval between 0 and 10 into sub-intervals of length one, then the probability of selecting a number between 0 and 2 will be the sum of the probabilities of selecting a number between 0 and 1, and between 1 and 2, i.e., 1/5. Intuitively, the longer

[7] A Lebesgue measure spreads probability over the points in a continuous manner (i.e., the difference in probability at two points can be made arbitrarily small by choosing two sufficiently close points). But an n-dimensional space has uncountably many points (n-tuples of real numbers). Therefore, if one of the points were assigned positive probability, there would be uncountably many other points close enough to the first one to also obtain positive probability. As a consequence, the probability of the union of those points (and, *a fortiori*, the probability of the whole space), would be infinitely large. This would contravene *Axiom 2* (the probability of the universal event is finite, namely 1). Therefore, a Lebesgue measure must assign every point in a space the same zero probability.

[8] The notion of randomness is far from clear; it is usually vaguely equated with chance, with a lack of order or certainty. Although strict definitions of randomness do exist, for our purposes the intuitive understanding will usually be sufficient.

an interval, the likelier it is that, other things being equal, it will contain the randomly selected number, since the limiting case (0 to 10) *must* contain that number by stipulation.

Densities and distributions

However, if after a few random trials the numbers picked out tended to centre around 6.4 and not around 2.5, it is tempting to say that 6.4 is more probable than 2.5. An interval of length one that includes 6.4 (say, 6 to 7) has a greater "density" of probabilities of respective outcomes than a same-length interval that includes 2.5 (say, 2 to 3). A formal expression of this intuition is the notion of *probability density* that is a generalisation of the notion of probability. A probability density is a function defined on points in an interval, whose values are non-negative. The integral of the probability density function on the points belonging to a continuous interval gives the probability of the *random variable* falling within that interval:

$$p(a<x<b) = \int_a^b Pxdx$$

where p is probability and P - probability density.

A random variable is such an important notion in probability theory and its applications, it is worthwhile giving a formal definition:

> A random variable is a quantity X capable, depending on the actual state of affairs, of taking on different numerical values, and having a definite probability of being found in any open or closed interval of real numbers (Howson and Urbach 1994, 27).

Also an important notion is that of a *distribution*. A distribution is a function assigning a probability (or probability density) to each value of a random variable. A random variable with a discrete range has probabilities distributed over it, while a random variable with a continuous range - probability densities. Once one knows a distribution, one can also calculate the probability of any subset of the set having the property in question. With a

discrete distribution, this is done by summation. And since integration is the continuous analogue of the operation of summation, in subsequent exposition many results involving summation in the discrete case preserve their validity in the continuous case provided that summation is substituted with integration.

The very distinction between discrete and continuous cases is far from rigid, at least in applications. Very fine discrete partitions (i.e., those where the number of elements of the field F is large, which includes all the countably infinite cases) are "quasi-continuous" in the sense that there exist continuous distributions whose properties they approximate the better the finer is F. For instance, the so-called "normal" distribution, which is continuous, is a good approximation of many discrete distributions, provided the number of elements in the latter is large enough. This can be very useful, since for a large number of elements, approximating a discrete distribution by a continuous function and finding its integral on the relevant interval may be easier than summing up all the probabilities in the original discrete distribution.

Another very important theorem of the probability calculus is that unconditional probability can be presented as a weighted sum of conditional probabilities. Assume A is the event in question. Let us divide the probability space into a set of mutually exclusive and jointly exhaustive events B_i such that $i=1,\ldots, n,\ldots$; $B_1 \cup \ldots B_n \cup \ldots = F$; $B_i \cap B_j = 0$ for all i, j, $i \neq j$. Then $p(A)=\sum p(A|B_i)*p(B_i)$, where $p(B_i)>0$ for all i. This result is referred to as the Law of Total Probability and is extremely useful in calculation.

CLASSICAL INTERPRETATION

Independence

Having established the similarities and differences between the ways in which probability can be assigned in discrete *vs.* continuous cases, we shall look at one further result that follows from the axioms of probability, first formally presented in the 17th century by

Thomas Bayes, and subsequently proved by Laplace. The Bayes Theorem (in its modern form)[9] states:

$$p(x|y)=p(x)*p(y|x)/p(y),$$

i.e., the probability of x conditional on y equals the unconditional ("prior") probability of x multiplied by the probability of y conditional on x (the "likelihood" in the technical sense of the term) and divided by the prior probability of y.

Before looking at an example, let us become clear about the kinds of cases that Bayes, Laplace and other fathers of probability theory had in mind. Since probability theory arose as a reflection on the games of chance, it was characterised by two features. First, it was conceived explicitly as a way of *measuring chances*. Second, it was concerned primarily with discrete distributions, moreover, a particular class of them - those in which every "atomic" state of affairs is *equally probable*. If the equiprobability assumption is satisfied, then measuring probability is very simple. The probability of an event equals the number of all those (equiprobable) states of affairs that are *favourable* to the event, divided by the *total* number of (equiprobable) states of affairs. Let us take the standard 52-pack of playing cards (excluding the jokers). The picking of each of the 52 cards is a separate state of affairs. There are 4 kings in the pack, so the probability of randomly picking a king equals 4 (the number of favourable states of affairs) divided by 52 (the number of all possible states of affairs), which equals 1/13. One can verify that this interpretation satisfies all the axioms of probability theory, including additivity. Indeed, the probability of picking either a jack or a king (8/52=2/13) equals the sum of probabilities of picking a jack and picking a king (2*1/13=2/13).

Similarly, in a fair die (i.e., one where each of the six faces is equiprobable), the probability of each face coming up is 1/6, and the probabilities of "composite" events can be found with the help of the axioms and theorems of probability calculus. Let us verify the Bayes

[9] Neither Bayes nor Laplace had notation for conditional probability.

theorem. According to it, the probability of a die rolling a number greater than 4, provided the die rolls an even number, equals the unconditional probability of the die rolling a number greater than 4 (i.e., 5 or 6), which for a fair die is 1/3, times the probability of an even number conditional on the die rolling a number greater than 4, which is 1/2, divided by the unconditional probability of the die rolling an even number, which is also 1/2. The result is 1/3, which makes sense: among the three scenarios that are permitted by the condition of rolling an even number (2, 4, and 6), precisely one is favourable to the desideratum of obtaining a number greater than 4 (in this case, only 6 will do).

Finally, let us demonstrate that, in the framework of the classical theory, the equiprobability of the sequences of trials that form a probability space implies the independence of those trials.[10] Take a sequence of n tosses of a fair coin. There are 2^n possible sequences of heads and tails of length n, all of which are equally probable. Let us denote the outcome of the ith toss ($1 \leq i \leq n$) "X_i". Then, by the classical definition of probability, prob(X_i=h) equals the number of sequences in which the ith toss yields heads, divided by the total number of sequences, namely 2^n. For simplicity, take n=2. Then 2^n=4. Of the four sequences, exactly two (namely, <h,h> and <h,t>) satisfy the proposition X_1=h. Therefore, prob(X_1=h)=2/4=0.5. The same holds for X_2=h, X_1=t and X_2=t. Analogously, inspecting the four possible sequences yields prob(X_1=h & X_2=h)=0.25. But, as we have seen, prob(X_1=h)*prob(X_2=h)=0.5*0.5=0.25. Combining the two equations yields

$$\text{prob}(X_1=\text{h} \;\&\; X_2=\text{h})=\text{prob}(X_1=\text{h})*\text{prob}(X_2=\text{h}),$$

hence, by the definition of independence, X_1 and X_2 are independent. This reasoning generalises to any finite n.

Counting favourable outcomes

Being so powerful and elegant in its original (and, it is easy to see, very limited) domain, the classical probability theory greatly stimulated the development of combinatorics - the

[10] The proof follows Howson and Urbach (1994), 42.

branch of mathematics that deals with *counting* separate states of affairs. If there are n objects, and out of those n we want to select m different ones, than the number of ways in which such a selection can be made is given by

$$^{n}Cm = n!/m!(n\text{-}m)!$$

If, in addition, each object has the same chance of being selected, we can derive some interesting consequences (such as Laplace's Rule of Succession, to be discussed shortly).

Apart from the games of chance, a favourite model was that of an urn with a certain number of balls in it, out of which out could select one ball at a time. One can select either "with replacement" (having retrieved a ball, one then puts it back, so that both the number and composition of balls remain unchanged), or "without replacement" (the selected ball remains outside, hence the number of balls decreases and their composition changes). Let there be an urn containing 5 black and 10 white balls. Suppose we want to retrieve a white ball, then put it back and take out a black one. Since we replace the first ball, the second trial is independent of the first one, and the probability of the desired sequence is simply the product of the probabilities of the two trials: 2/3*1/3=2/9=10/45 (about 22%). If we do not replace the first ball, then the probability of the desired sequence is 2/3*5/14=10/42 (about 24%) - slightly better odds than before. Likewise, if each selection includes more than one ball at once, we can calculate the number of favourable possibilities and their ratio to the total number of possibilities. Say, we want to retrieve three balls at once. What is the probability that one of them will be white and two - black? Combinatorics tells us that the number of ways in which two black balls can be selected out of five is 10, and the number of ways in which one white ball can be selected out of ten is 10, so the total number of favourable outcomes is 10*10=100. The total number of ways in which 3 balls (not necessarily of the required colours) can be selected out of fifteen is 455. Hence, the probability of the desired selection is 100/455=20/91. If we want to obtain this same selection twice in a row, with replacement, the probability drops to $(20/91)^2$=400/8,281, which is less than 5%. So long as a domain can be feasibly modelled on an urn, the probability of anything and everything happening in it is measurable!

One example of the power of combinatorics is Laplace's famous derivation of the Rule of Succession which, purportedly, proved that, *pace* the Sceptics, we can be virtually certain that the Sun will rise tomorrow. Indeed, the Rule of Succession states that if an event has recurred m times out of the possible n, its probability of recurring for the m+1st time equals $(m+1)/(n+2)$. Given that there are no records of the Sun ever failing to rise, $m=n$, and therefore the probability of its rising tomorrow is $(m+1)/(n+2)$ which, for the extremely large observed m, virtually equals 1.

Unfortunately, we cannot afford to be as sanguine about the Sun rising tomorrow as Laplace would have us. Indeed, a crucial assumption in his derivation is that the Sun's success or failure to rise on n consecutive days are independent, which does not seem plausible. If it had not risen today, I would have suspected that something had happened to it that would prevent it from rising again any time soon.

Symmetry and ignorance

The restrictive character of the classical theory can be further illustrated by looking more closely at the equiprobability assumption. What grounds do we ever have for assuming that two options have exactly the same chance of occurring? One possible line of argument is from *symmetry*. If a die is a perfect cube, then all its faces are indistinguishable save for the different numbers of dots on them. *Ergo*, each of the faces must have the same chance of coming up. The second line of argument concerns those cases where we cannot claim perfect symmetry say, in betting on the outcome of a horse race. If we do not know anything about the horses except their names (or numbers), we cannot go as far as to claim that each has the same objective chance of success, but our *ignorance* of any relevant differences compels us to assign them all an equal probability of winning. Hence, Classical probability is often characterised as "measure of ignorance". Its central expression is the famous Principle of Indifference: if there is no information favouring one of the possible states of affairs, then they must all be equiprobable.

This is a stronger claim than the Symmetry Argument: given a system that is visibly asymmetrical, we should still view the different outcomes as equiprobable unless we have reasons to think otherwise.[11] (That is why the principle is often referred to as the Principle of Insufficient Reason.) Take a coin that we *know* to be biased but the bias is not reflected in its appearance. Prior to experimentation, we should view heads just as probable as tails, although the objective chance of one is greater than the other. This conundrum was to develop later into an explicit split between epistemic *vs.* physicalist interpretations of probability, of which later.

At least, in the biased coin case the Principle of Indifference can be defended on the grounds that equal lack of information about the possible outcomes renders choosing one as rational as choosing the other. Suppose one is a goalkeeper trying to save a penalty kick. It is known that the ball will be struck strongly into one of the two bottom corners, depending on the preference of the taker. If that preference is unknown, the keeper has as much chance, *on average*, to save the kick by diving into one corner as into the other, and therefore he would be equally rational in diving to the left as to the right (unlike a case where the taker's preference is known). The trouble with the Principle of Indifference is not that it may give equal probability to outcomes whose chances are, in fact, different. It is that often there is no unique way of delineating the outcomes that are supposed to be equiprobable.

Take a library catalogued by the author and placed, in the alphabetical order, in stacks numbered 1-10. Suppose we want to find the source of a particular *bon mot*, not knowing the author. The letter B happens to completely take up stack number 3, and no part of another. What is the probability of finding the relevant book in stack 3? Since there are 10 stacks altogether, and we have no particular preference for any one, the Principle of

[11] The visible asymmetry is not by itself enough, since we must have a theory about how exactly that asymmetry translates into different likelihoods of different outcomes. OK, today it is warm and dry and tomorrow it will be cold and wet. Possibly, one of these types of weather is more conducive to asthma attacks than the other, but in the absence of a substantive theory of asthma we cannot say which, let alone assign non-equal probabilities to little Johnny suffering an attack on one of these days as opposed to the other. An objectivist would still say that the two probabilities are objectively unequal, but his assertion will ring hollow until he provides such a substantive theory.

Indifference gives stack 3 probability 1/10. However, look at it this way. Since we don't know the author of the quotation, it could be any one represented in the library. Assuming that there are equal numbers of authors under each letter, the letter B has the same probability of yielding the required quotation as any other, that is, 1/26. Since all the authors starting with 'B' are located in stack 3, stack 3 receives probability 1/26. So, which is the correct probability, 1/10 or 1/26? The classical approach had no convincing answer.[12]

The principle flounders similarly wherever there is more than one "legitimate", or "natural", or "intuitive" way of partitioning the outcome space. Unsurprisingly, it completely breaks down in continuous cases where any partition has to be *imposed* on the system rather than *found* in it.

Take a perfectly round lake inhabited by some fish that emerge to the surface randomly. John, Brian and Mary are observing the lake from the same point on its shore, having been assigned three mutually exclusive and jointly exhaustive segments so that each is responsible for the same width of the horizon (i.e., 60°). The Principle of Indifference will give ambiguous predictions about the probability of Brian (who is in the middle) spotting a fish first. Since he has exactly 1/3 of their total field of observation, the probability must also be 1/3. However, the surface area of the lake that corresponds to his field of observation, is greater than 1/3 (in fact, it is approximately 7/12, more than half). In general, if a certain function describes a continuous domain, any strictly increasing transformation of the original function will describe it just as well, and what was an "equal" division of the domain under the original description, may be a very "disproportionate" one under the resultant description. True, it may be possible to argue, on a case by case basis, that a particular description is "more natural" than all the others (in this example, surface area seems more germane than observation field width). However, choosing such a preferred description requires a substantive argument about the nature of the domain, its causal set-up, rather than pleading equal ignorance about the possible outcomes.

[12] It may be objected that the example is not a very good one, as there are many more authors under certain letters than under others. However, a rational being with only limited familiarity with Indo-European languages and the Latin alphabet would not be aware of that.

The classical approach quite correctly saw probability as a measure of possibility. What "objectively" has a greater chance of occurring, should be "subjectively" viewed as more probable. The trouble was, very little insight was offered into objective chance[13]. The symmetry argument is largely negative. What if symmetry does not hold? Once one goes beyond "perfectly symmetrical" dice, "perfectly shuffled" packs of cards, and "perfectly identical" balls "perfectly randomly" retrieved from an urn, there was nothing that could be said about the nature of things that underlied their "random" behaviour. Since the 1920's, a number of very different approaches were developed in order to account for the cases where the assumptions of the Classical interpretation could not be trusted to hold.

PHYSICALIST INTERPRETATION

Von Mises: empiricism

Collectives

What I refer to as the Physicalist Interpretation, is in fact, a collection of quite distinct approaches linked by the common idea that probability is a measure of objective chance. As such, probability judgements must be founded in the behaviour of the system. Just as the two sides of a fair coin coming up are equiprobable because the coin is fair, i.e., its two sides have an equal natural tendency to come up when tossed in a "random" fashion, in a biased coin, one of the sides will come up more frequently because the mass distribution within the coin is uneven, and it has a natural tendency to land on the heavier side. Whereas most of the fathers of classical probability shared Laplace's determinism which left no room for objective chance and therefore confined probability to measuring the degree of our ignorance, the founders of the physicalist interpretation such as Richard Von Mises conceived of probability as an empirical science on a par with physics that used

[13] This is not surprising given the determinism so prevalent in 17-19 centuries and famously defined by Laplace: if one could know all the initial conditions of a physical system (including the world as a whole) and all the laws governing it, and possess unlimited powers of computation, then one could predict the exact state of the system at any point in time. Uncertainty has no place in the deterministic *ontology* (despite its very prominent place in Laplace's *epistemology*), therefore objective chance does not exist. Probability cannot be anything other than a measure of ignorance.

experimental techniques to discover the true properties of physical systems, from dice and coins to decaying atoms to social institutions.

Any system whose behaviour is capable of randomly exhibiting alternative outcomes, and therefore can be described in probabilistic terms, was dubbed my Von Mises a "collective". The term 'collective'

> ...denotes a sequence of uniform events or processes which differ by certain observable attributes, say colours, numbers, or anything else. (...) All the throws of dice made in the course of a game form a collective wherein the attribute of the single event is the number of points thrown. Again, all the molecules in a given volume of gas may be considered as a collective, and the attribute of a single molecule might be its velocity. A further example of a collective is the whole class of insured men and women whose ages at death have been registered by an insurance office. The principle that underlies the whole of our treatment of the probability problem is that a collective must exist before we begin to speak of probability. The definition of probability which we shall give is only concerned with 'the probability of encountering a certain attribute in a given collective' (Von Mises 1957, 12).

Elsewhere, Von Mises points out that, in the strict sense, a collective is a *finite* sequence. However, for a large number of phenomena, their behaviour can be successfully represented *as if* they formed an infinite sequence. If the number of elements in a sequence is potentially infinite, they can be viewed as forming a *bona fide* collective, and therefore their behaviour can be viewed as random.

Randomness I

What is "randomness"? For Laplace, it was synonymous with ignorance. When a coin is tossed, it is destined to land on a particular side. We call the outcomes of experiments like tossing a coin "random" only because we do not fully know the initial conditions and/or cannot calculate the trajectory. Von Mises viewed randomness more literally. Before we discuss the many problems associated with this notion, let us state in a nutshell how randomness was viewed by Von Mises.

An event is random if, in repeated trials, the outcomes of the previous trials do not allow us to predict the outcome of a subsequent one. Say, the observed sequence of five tosses is h-h-t-h-h. Despite the apparent surfeit of heads (assuming the coin is fair), we cannot conclude that the sixth trial will yield tails "to even the balance". If such inferences were correct, one would be able to wait for a "deficit" of tails and then bet on the "under-represented" outcome. As many a disappointed gambler will testify, this reasoning does not work empirically. That is because a series of coin tosses is a collective, and as such, subject to **The Law of Excluded Gambling Systems**:

> *A sequence of trials is random if it is not possible to devise a system that would take the results of the observed trials as input and produce a more-than-average proportion of correct guesses about the outcomes of future trials.*

The second essential feature of a collective is that it obeys **The Law of Stability of Collectives**:

> *As the length of a sequence of trials increases, the relative frequencies of different outcomes converge to fixed limits.*

In a short sequence, virtually any technically possible proportion of heads and tails is fairly probable (e.g., in a sequence of six tosses, the probability of 2/3 of the trials yielding heads equals 23%, whereas the 'more equitable' proportion of 1/2 heads in six tosses is only slightly more probable, at 31%). As the length of a sequence of tosses increases, the relative frequencies converge to certain limits. Take the following result of randomly tossing a fair coin 16 times:

h-t-t-t-h-t-h-h-t-t-t-t-h-h-t-h

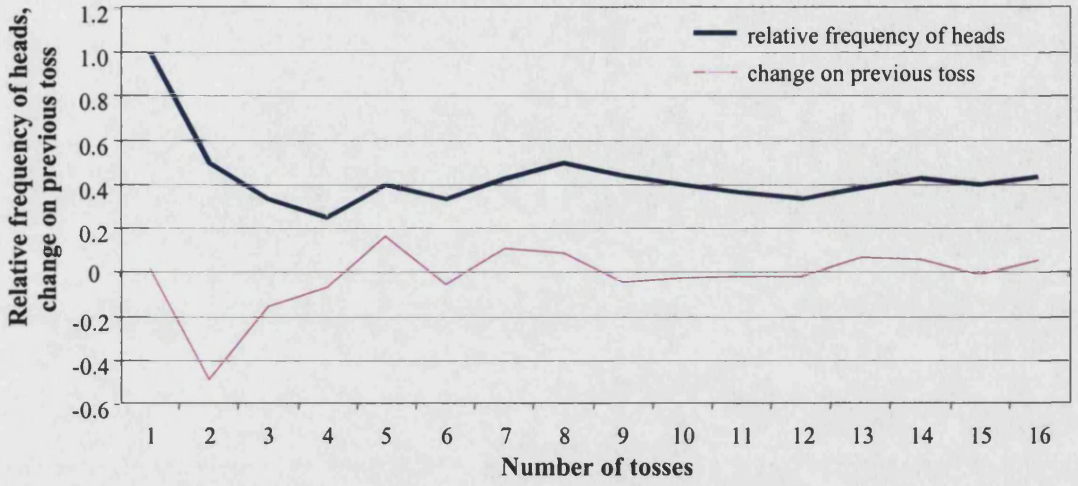
Let us then summarise the relative frequencies of heads and tails for different n, and the on-step differences between them. The result is presented in table form, and for heads - also as a graph.

Table 1: Relative frequencies of heads and tails, and their changes on the previous toss

n	r_h	change on previous n	r_t	change on previous n
1	1.00	n/a	0.00	n/a
2	0.50	-0.50	0.50	0.50
3	0.33	-0.17	0.67	0.17
4	0.25	-0.08	0.75	0.08
5	0.40	0.15	0.60	-0.15
6	0.33	-0.07	0.67	0.07
7	0.43	0.10	0.57	-0.10
8	0.50	0.07	0.50	-0.07
9	0.44	-0.06	0.56	0.06
10	0.40	-0.04	0.60	0.04
11	0.36	-0.04	0.64	0.04
12	0.33	-0.03	0.67	0.03
13	0.39	0.05	0.62	-0.05
14	0.43	0.04	0.57	-0.04
15	0.40	-0.03	0.60	0.03
16	0.44	0.04	0.56	-0.04

Note: Relative frequencies and changes have been rounded to the second decimal point.

Figure 1: Relative frequency of heads and its change on the previous toss



Convergence

Initially, we observe fairly wide swings in relative frequencies. As n increases, however, the swings become progressively smaller, in the limit vanishing altogether. What are the limits to which the relative frequencies converge?

First of all, let us see how probable is the heads/tails sequence reported above. It is easy to show that the probability of a sequence depends on its length. For n=2, we have four possible sequences: h-h, h-t, t-h, and t-t. Assuming the coin is fair, each sequence is equiprobable,[14] with probability ¼. For n=4, there are 16 equiprobable sequences, and for n=16, there are 65,536. Hence, the probability of our sequence is 1/65,536 (less than 0.0015%). However, the probability of *any* sequence with the observed ratio of heads (r_h =7/16) is a lot higher – just over 17% (which is just a bit less probable than a sequence with r_h= r_t=8/16, at 19.3%). The probability that in an n-long sequence the relative frequencies are within 1/16th of ½, is therefore more than half – 53.5%. So, both r_h and r_t converge to ½, which happens to be the probability of heads and tails in a fair coin. Can we assume,

[14] That is, the expected gain of betting on one of these sequences does not depend on which one we chose.

from the convergence of the relative frequency of an outcome in repeated trials to x, that x is indeed the probability of that outcome? In this example, does the convergence to ½ indicate that ½ is the probability of heads and tails and the coin is therefore fair?

Intuitively, the more times we toss the coin, the more reliable our judgement of the probabilities of heads and tails will be. In the extreme case of just one toss, the outcome of tails does not tell us virtually anything (in particular, it does not indicate any bias), since one or the other side has to come up, be the coin fair or biased. On the other hand, "in the long run", any significant deviation from the 50/50 proportion almost certainly indicates a bias (assuming that the tosses are independent). Let us see how the probability of a certain relative frequency in a sequence depends on its length. Using the usual combinatorial arguments, if the tosses are independent and in each one, heads and tails have equals probability (1/2), then the probabilities of sequences with the relative frequencies of heads (r_h) and tails (r_t) being {1/2; 1/2} and {1/4; 3/4}, are as follows:

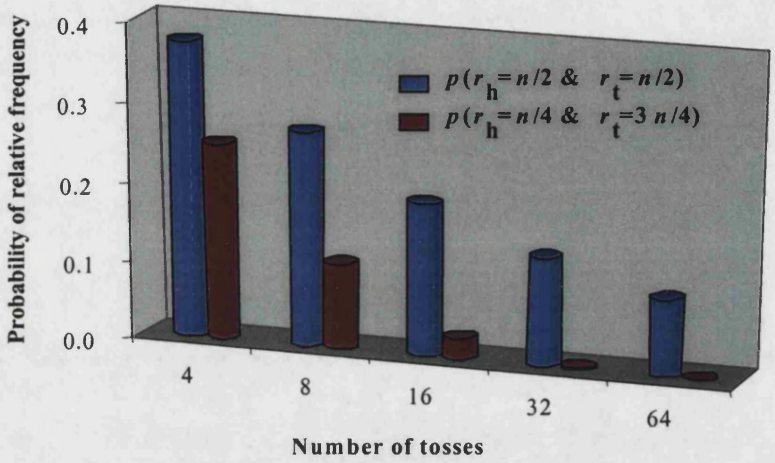
Table 2: Probability of a given relative frequency vs. the number of tosses

n	$p(r_h=1/2 \ \& \ r_t=1/2)$	$p(r_h=1/4 \ \& \ r_t=3/4)$
4	0.375	0.25
8	0.273	0.11
16	0.196	0.027
32	0.140	0.00245
64	0.100	0.000265

As n increases, so does the number of different relative frequencies of heads and tails possible in each sequence; therefore the total probability of a given frequency being realised, which is equal to one, gets distributed "more thinly". Although the probability of the proportion of heads in a sequence being equal to the probability of heads in each given toss decreases with the length of the sequence, it does so much more slowly than the probability of a "rogue" proportion, such as 1/4. If one pictures the probability distribution of different relative frequencies by plotting the frequencies on the horizontal axis and their

probabilities - on the vertical, two things become obvious. First, the graph becomes smoother (since the number of elements in the partition increases). Continuous approximations of discrete distributions mentioned above are based on this fact. Second, the peak corresponding to the "true" frequency gets relatively steeper.

Figure 2: Probability of a given relative frequency vs. the number of tosses



This can be, and is, used to estimate the "true" probabilities of various outcomes. If no other salient information is available, then the best estimate of the true probability is, subject to certain conditions, that which corresponds to the peak in the probability distribution of the relevant frequencies.[15] In the figure above, $r_h=r_t=1/2$ is much the likelier to be the true relative frequency, since its probability calculated from observing the actual sequence of the first n tosses, decreases considerably more slowly than that of the alternative.

The possibility of estimating the true probabilities of outcomes is formally reflected in the so-called Laws of Large Numbers which relate the relative frequencies of outcomes to their

[15] If a sample distribution has only one peak, and is symmetrical in relation to it, then the peak corresponds to the *mean* of the sample distribution. The mean value of a distribution is the probability-weighted average of all its values. Further, according to a theorem of mathematical statistics, the mean of a sample is an unbiased estimate of the (unknown) mean of the population from which that sample has been drawn.

probabilities.[16] The Weak Law of Large Numbers states that as n increases, the probability of the relative frequency of an outcome (in a *finite* sequence) being arbitrarily close to its true probability, converges to one. The Strong Law of Large Numbers goes further in saying that in an *infinite* sequence, the relative frequency of an outcome with probability one converges to a limit which, in turn, equals the true probability of the outcome. There is only a small step from here to saying that with probability one, relative frequency *equals* the true probability "in the limit". (Of course, there is still the problem of how the large finite relates to the infinite, and it is this problem that renders the application of the Laws of Large Numbers to real-life phenomena less than straightforward.)

Von Mises was impressed by the ubiquity of phenomena exhibiting seemingly random behaviour that are subject to the Laws of Large Numbers. According to the Laws, measuring relative frequencies is, in the limit, the same as measuring the underlying probabilities. Being a positivist, however, Von Mises had little taste for multiplying entities. If relative frequencies are the only empirical measures of probabilities, they *are* probabilities, just as for most of the logical positivists the meaning of 'temperature' was fully given by the set of procedures in which temperature was measured.

But, as mentioned, relative frequency accurately measures probability only in the limit, i.e., when n is infinite. Even then, the two coincide only 'with probability one' rather than 'in all cases'[17] (although many find it convenient to ignore this caveat and define one through the other, just as Von Mises suggested). However, the first point had to be addressed, and Von Mises' definition of probability (which, we remember, was meant to apply only to 'collectives') required a particular interpretation of a collective as an infinite sequence of

[16] Provided that the trials are independent and the probability distribution of the outcomes is the same in each trial. In such trials, the types of outcomes are referred to as 'independent, identically distributed random variables'.

[17] 'Probability one' does not mean 'always', just as 'probability zero' does not mean 'never'. With probability one, if we were to repeat an infinite sequence of trials infinitely often, we might still obtain a relative frequency that was different from the true probability, but only in a finite number of cases. In other words, the cardinality of the class of exceptions is lower than the cardinality of the class of non-exceptions.

outcomes of an infinitely repeatable experiment. Since all actual sequences are finite, the emphasis had to be on *potential* repeatability.

Randomness II

Let us state Von Mises' axiom of randomness more formally. We shall call any method whereby we chose an infinite sub-sequence out of the whole sequence, a "selection procedure". An *admissible* selection procedure is such that it is based only on the positions of the elements (trials) in the whole sequence, plus the outcomes of *prior* trials. For instance, we might want to select only the odd, or even, or primary-numbered trials. Or we could select every trial that follows the sequence "h-h-t". Or the odd-numbered trials that follow such a sequence. On the contrary, a sub-sequence based, partially or fully, on the outcomes of the *future* trials, would not be admissible. The axiom of randomness then stipulates that in all infinite sub-sequences generated by an admissible selection procedure, the limiting frequencies of outcomes are the same as in the original collective. So, if rolling a die constitutes a collective, we cannot obtain the limiting frequency of a particular face different from that frequency in all the trials (for a fair die, 1/6). Therefore, neither can we expect a higher-than-average win in betting on the face of a die by betting in every third game, or by using a more elaborate formula. It is clear why only some selection procedures are admissible: we would have dearly loved to bet on a particular outcome only in those trials in which that outcome obtains, but that is not practically possible.

If, empirically, a sequence of trials (1) exhibits converging relative frequencies (The Law of Stability) and (2) is immune to any permissible selections (The Law of Excluded Gambling Systems), it satisfies the Axioms of Convergence and Randomness and is therefore subject to the probability calculus. If either of the two empirical conditions fails, then no matter how much a sequence of repeated trials looks like a collective, it is not one.

The question then arises, are these conditions too onerous? Are such random sequences, or "collectives", ubiquitous? If they were not, it would be scandalous for the frequentist

interpretation since it would render the notion of probability vacuous. The question is more than simply one of existence. It is possible to provide any number of sequences such as 00000000,...000,..., where any place selection procedure yields a sub-sequence with exactly the same relative frequencies. Unfortunately, such trivial instances are insufficient to prove the applicability of Von Mises' notion of probability to real phenomena. In order for it to be applicable, there must be sufficiently many non-trivial sequences that obey Von Mises' axioms.

As it happens, it was subsequently proven that random sequences do exist in sufficient quantity. The mathematician Abraham Wald demonstrated in 1937 that if a set of selection procedures is denumerable (i.e., finite or countably infinite), there exists an uncountable set of sequences whose relative frequencies are invariant with respect to the selection procedures. In other words, to each denumerable set of admissible selection procedures there corresponds a non-denumerable set of bona fide collectives.

Of course, what remained to be proven was Wald's denumerability assumption. Luckily, the proof was provided very shortly by the logician Alonzo Church in his (1940). He developed the notion of a *recursive function.*[18] Such a function is interesting in that it can algorithmically perform calculations of arbitrary length. Church showed that the class of recursive functions is denumerable. Consider the set of formulae that determine each given function: the length of a formula is finite, and there are only finitely many formulae of each given length.[19] Since there are at most as many recursive functions as formulae, the former are denumerable. Thus Wald's condition was shown to hold and therefore collectives - to exist in sufficient number.

[18] At or around the same time, a number of mathematicians such as Turing, Markov, Gödel, Herbrand and Post introduced similar notions. However, it was Church whose work was directly linked with the assumptions of Wald's theorem.

[19] This inference assumes that each recursive function is defined by a set of formulae. However, the assumption had previously been proven by Gödel and Herbrand.

A separate explication of randomness was later provided by Kolmogorov (1965): a sequence is random if the shortest algorithm that calculates its sub-sequences is at least as long as the sub-sequences themselves. Kolmogorov's definition based on the theory of computability is stronger than Wald's, as it also covers finite sequences.

Popper: conjectures and refutations

Anti-inductivism

As previously discussed, Von Mises' understanding of probability involves *potentially* rather than *actually* repeated trials. However, a positivist is on shaky ground with potentiality, since it requires a non-verifiable internal cause guaranteeing an infinity of repetitions. Potentiality of repetitions is implicit in any scientific law, so this is a general objection against positivist philosophy of science, and not probability-specific. The late Wittgenstein strongly criticised the habit of uncritically assuming that any particular sequence would continue *ad infinitum*.[20] Hence, such an assumption can only be a falsifiable (and probably false) hypothesis. This, in a nutshell, is the position arrived at by Karl Popper (who, of course, was not familiar with Wittgenstein's post-*Tractatus* musings). Von Mises published his account of probability in the 1920's; less than 10 years later Popper turned Von Mises' idea on its head. Those acrobatics resulted from more general methodological arguments.

Popper based his methodology on the fact that while no finite number of singular instances entails a general hypothesis, even one counter-example can refute it. Therefore, the *modus operandi* of empirical science cannot be observing the facts and inferring general hypotheses from them. Instead, scientists come up with hypotheses and check how they fare against the sea of potential counter-examples (or even "attempt a vigorous refutation"). Once a hypothesis is refuted, we make up a new one such that it survives all the existing

[20] Take a certain experimental set-up yielding 10 zeros in a row. It could be a manifestation of a law of nature mandating a zero in each repetition, or it could exemplify an alternative law mandating 10 zeros in a row followed by a one, followed by another 10 zeros, and so on. Later, Wittgenstein's idea of "bent rules" was developed by Nelson Goodman in his famous "grue/bleen" example. Unlikely as they look, "bent" laws of nature are not unheard of in science, and certainly cannot be ruled out without argument.

counter-examples to its predecessors, while at the same time exposing itself to some new ones, and so it goes on and on. In the process, we end up with better theories than before. Hypotheses about true probabilities are, for Popper, on a par with all the other scientific hypotheses. Yes, relative frequency *is* inextricably related to probability, just as Von Mises said. But we do not infer true probabilities from observable sequences; we *hypothesise* the former and test them with the latter.

That Popper's suggestion immediately runs into a lethal methodological difficulty, should be clear. Statements about limiting frequencies are unfalsifiable as well as unverifiable. Some finite sequences are extremely *unlikely* (i.e., have *low* conditional probability) on certain assumptions about the limiting frequencies 'underlying' them, but they are never *inconsistent* with such assumptions (which would require *zero* conditional probability).[21] Therefore, there is no methodological advantage to adopting Popper's rather than Von Mises' view.

Propensity

But let us, for the sake of an argument, grant Popper that probability is *not* limiting relative frequency. Then what is it? According to Popper, it is the 'propensity' of an experimental set-up to produce certain relative frequencies.[22] A fair coin, tossed in the usual manner onto a hard surface, has a 0.5 probability of coming up heads because its internal structure in relation to the rest of the experimental set-up is such that it tends with equal strength to land on either side. A propensity in Popper's understanding is a causal tendency, and in fact he defined causation itself in terms of propensities: A *causes* B **iff** A has a 100% propensity of producing B. As Donald Gillies puts it,

[21] I shall discuss shortly Popper's (unsuccessful) attempt to get around this difficulty.

[22] More on the notion of propensity, and its role in both Popper's and, implicitly, Von Mises' theories, see below.

> ...a large dose of cyanide will definitely cause death. A suitably small dose of cyanide might only give rise to a propensity of 0.6 of dying. So propensity appears to be a kind of weakened form of causality (Gillies forthcoming, 1).

However, despite its initial appeal, this idea is faulty in at least two ways. Firstly, there is the so-called Humphreys' paradox. In Gillies' exposition, the problem is as follows:

> Causes have a definite direction in time. So if A causes B and A occurs before B, then B does not cause A. Apart from a few speculations in theoretical physics, it is universally conceded that causes do not operate backward in time. The situation is very different with probabilities [and therefore with propensities – **M.M.**]. For events A, B, we usually have that if P(A | B) is defined, then so is P(B | A). Probabilities have a symmetry where causes are asymmetrical. It thus seems that propensity cannot after all be a generalisation of cause (Gillies forthcoming, 9).

The second difficulty with Popper's view of propensities becomes apparent when comparing the models known as 'Bayesian networks' and 'causal networks'. Each of those is a 'directed acyclical graph' (DAG) with the appropriate relation defined between certain nodes of the graph. In the former case, the relation between the nodes is that of conditional probability, while in the latter case it is that of being a possible cause. The two types of networks are often identified with each other, as indeed they would need to be in order for probability to be isomorphic with causality. However, as Gillies shows on a simple example, such an identity does not hold in the general case as many Bayesian networks are not causal networks, and vice versa. This indicates that propensities cannot be explicated through the concept of a cause.

Be that as it may, on any account propensity is a dispositional property, i.e., its presence does not by itself guarantee any particular manifestations. It is Popper's contention that this understanding of probability applies to those cases where there is no possibility of a sequence of repeated trials, let alone of an infinite sequence. According to Popper, this advantageously differentiates his view from Von Mises'. Let us separate two issues that can sometimes be confused:

(1) Can the notion of probability be meaningfully applied to a finite sequence? Both Popper and Von Mises answer in the affirmative, since, as we have seen, the latter also held a dispositional view. There is little difference between the two on this count.

(2) Can the notion of probability be meaningfully applied to a *singular event*? Once again, this has to be disambiguated, focusing first on 'singular' and then on 'event':

(2.1) Can the notion of probability be meaningfully applied to a sequence of length one? For Von Mises, the affirmative answer follows from his similarly positive answer to (1): there is no principled difference between two finite numbers, and a sequence of length one is merely one of many sequences of different finite lengths. As long as a singular event can be viewed as an instance of a potential infinity of repetitions, a probability can be assigned to it. So, even if we flip a coin just once and then destroy it, it is perfectly meaningful to speak of the probability of heads being 0.55: we claim to believe that *had* the coin been flipped infinitely many times, the relative frequency of heads would have eventually converged to 0.55.

We can go further and stipulate that the coin to be flipped, although identical to the previous one in all other respects, self-destructs after the first toss. There is no reason to alter our judgement of the probability of heads. Hence, the physical impossibility of an event being repeated more than once does not by itself preclude us from assigning it a Von Misean probability.

The other aspect of the question is (2.2) Can the notion of probability be meaningfully assigned to an *event*? It is here that Von Mises radically differs from Popper. According to the former, events are not the right entities to which to assign probability, be that singular or repeated events. The domain of the probability function is made of *attributes* of (potentially) repeatable events (e.g., the attribute of 'heads' in an experimental set-up that potentially generates an infinity of coin tosses), not of events themselves ('the coin lands heads up'). Once that is understood, 'singularity' becomes a non-issue.

Events vs. attributes

Once the real issue between Von Mises and Popper becomes clear, we can judge the comparative merits of their approaches. Popper's definition of probability is remarkably similar to Von Mises', except that it is now defined for events rather than attributes. However, this difference is big enough to warrant a closer look.

According to Popper, an event has probability x if it results from a set of repeatable conditions such that, were they to be repeated n times, the relative frequency of the event would converge to x as n tended to infinity. We are speaking strictly hypothetically; although Popper was unclear about what made a set of conditions repeatable, he was adamant that no actual repetitions were required.

So, the probability of an outcome is not the property of an infinite sequence of trials, but rather of the experimental set-up which produces those trials. This opens the possibility of defining probability for those events that do not happen very often, or even happen just once. Unlike Von Mises, Popper would be prepared to define probability of such "singular events" as a rocket being successfully launched, for which no actual sequence of repeated trials is available. *Had* we launched the same rocket in exactly the same conditions many times, the resulting proportion of successful launches would be equal to the probability in question.

The trouble with assigning probability to individual events (rather than to attributes in a sequence) is that any event can instantiate any number of different repeatable conditions. This is traditionally dubbed the problem of the choice of reference class. Howson and Urbach illustrate the problem with the following die-tossing example:

> Suppose that the face on which the die falls is uniquely determined by parameters $q_1, \ldots, q_k$. Suppose also that at the first throw of the fair die, these take values $q_{10}, \ldots, q_{k0}$, and that the outcome of that throw is a four. Consider the experiment E_0 consisting in throwing the die in such a way that the q_i take the values q_{i0}. Clearly, the relative frequency of a six in any sequence of repetitions of E_0 is 0, whereas the relative frequency of a six in a long sequence of repetitions of the experiment E,

> consisting simply of throwing the fair die in the usual way, is one sixth. Now the first throw of the fair die ... is an instance both of E_0 and of E. Hence the single-case probability of a six at the first throw of the die is both 0 and one sixth, and we have a contradiction ... (Howson and Urbach 1993, 340).

For Von Mises, on the contrary, the problem of reference class does not arise, since probability is applicable only to the attributes in an already well-defined sequence. One may therefore claim that Popper's expansion of the scope of the physicalist notion of probability carries too high a price. These differences aside, there is more common between Von Mises and Popper than between either of them and, say, Laplace and Bernoulli. When all is said and done, the frequentist and the propensity interpretations are not essentially different: they both view probability as the property of a physical system (an experimental set-up), and see relative frequency as the link between true probability and our knowledge of it. Unfortunately, on both counts the two interpretations run into serious difficulties.

Wither 'virtual frequencies'?

Let us expand on the former similarity between Von Mises' and Popper's approaches, namely the view that probability is a property not of infinite sequences themselves, but of the *experimental arrangements* that produce them. An experimental arrangement, as both philosophers saw it, is a set of repeatable conditions that yield alternative outcomes with certain frequencies.

Every experimental arrangement is liable to produce, if we repeat the experiment very often, a sequence with frequencies which depend upon this particular experimental arrangement. These virtual frequencies may be called probabilities. But since the probabilities turn out to depend upon the experimental arrangement, they may be looked upon as *propensities of this arrangement.* In Popper's words, *they characterise the disposition*, or *the propensity* of the experimental arrangement to give rise to certain characteristic frequencies *when the experiment is often repeated* (1957, 67).

However, some events have *no* unique set of repeatable conditions with which they are associated. We would like to be able to talk of the probability of the Sun going super-nova in the next million years, however the conditions associated with this event are not repeatable, even hypothetically. Even when they are, one cannot simply stipulate "make it all just as it was, and repeat the experiment". The conditions under which the experiment is performed must be stated explicitly, and no explicit description (being, speaking technically, denumerable) can fully specify the exact set-up, which is continuous. There will always be errors of measurement resulting in subtle differences in the set-up from one trial to the next, and who can guarantee that those differences, imperceptible as they may be to us, will not alter the probability of the outcome? To sum up, the device of repeatable conditions is too coarse to pin down a unique virtual frequency, which is a problem for both Popper and Von Mises.

Problems with additivity?

The second essential similarity between Von Mises' "strict" frequentism and Popper's "propensity" modification can also be seen as their shared weakness. The problem is that probability and limiting frequency fare differently with respect with additivity. Frequencies, be they limiting actual frequencies of Von Mises or virtual ones of Popper, are not always countably additive and on many important domains, not even additive. To borrow an example from Bas Van Fraassen (going back to Ronald N. Giere),

> Let $A(n)$ be an event that happens only on the nth day. Then the limit of the relevant frequency of the occurrence of $A(n)$ in the first $n+q$ days, as q goes to infinity, goes to zero. The sum of all these zeros, for n = 1, 2, 3,... equals zero again. But the union of the events $A(n)$ - that is, $A(1)$-or-$A(2)$-or-$A(3)$-...; symbolically $U\{A(n): n<N\}$ - has relative frequency 1. It is an event that happens every day (1980, 184).[23]

[23] To be fair to Von Mises, the event $U\{A(n): n<N\}$ is not the kind of event to which probability is assignable in his theory, therefore Van Fraassen's counter-example exposes not any inconsistency but, at most, a limitation of scope.

Examples like the above convinced Van Fraassen that if we want probability to be countably additive, it cannot be identified with limiting relative frequency. Before assessing the merit of Van Fraassen's position, let us look at another example that he uses.

Hans Reichenbach, another proponent of the frequentist interpretation, used the example of a machine gun randomly firing at a round target. His contention was that the probability of a region of the target being hit by a bullet is the relative area of that region in relation to the target as a whole. Van Fraassen modifies this example to make the bullets point-particles (with zero area). Consider the region hit by these points; its area is zero. So, the region hit with relative frequency *one* has a *zero* probability of being hit. The opposite holds for the complement of the hit region: it is never hit, so the relative frequency of being hit is zero, but its probability of being hit is one since it has the same area as the whole of the target. Van Fraassen thus defends the potency of such illustrations:

> Because I gave rather a fanciful example, this conclusion may not seem damaging. But it is; it merely brings out in a graphic fashion that there are probability spaces which are not isomorphic to any such 'relative frequency space' resulting from a long run of outcomes with a relative frequency function *relf*. And those spaces - the geometric probabilities - are just the ones that are mainly used in physics (Van Fraassen, 1980, 185).

Van Fraassen's contention is that the inapplicability of the Von Mises (frequentist) view of probability to geometrical probabilities invalidates any claim that view might have to representing modern scientific practice. This, however, is a *non-sequetur*. As Howson and Urbach point out, Van Fraassen's argument rests on his instrumentalist methodology and fails in a realist framework. For the instrumentalist, there is no 'fact of the matter' about a scientific theory, beyond it being entrenched in scientific practice. Hence, the conceptual apparatus of geometrical probability, which is thus entrenched, is viewed as the final arbiter of the acceptability of frequentism. The realist, on the contrary, views *both* geometrical and frequentist probabilities as equally imperfect but complementary ways of describing pre-theoretic reality:

> But a realist should no more feel forced to make a choice between limiting relative frequency and 'geometrical' probabilities than he should between a belief that a piece of matter is a discrete bundle of molecules and the adoption of a continuous function to describe its mass distribution. Continuous mass distributions are a fiction, but a useful one, for they offer a mathematically simple approximation to the true state of affairs. Likewise, 'geometrical' probability distributions are a sufficiently good approximation to appropriate relative frequency distributions to warrant their use as simple mathematical models of them (Howson and Urbach 1993, 328).

Probabilistic statements: 'metaphysical'?

In addition to the two difficulties shared by Von Mises' and Popper's versions of frequentism, Popper's has a further drawback, namely that his view of probability is incompatible with his own stated methodology. Popper saw one of the main advances of the propensity interpretation in that, supposedly, it corrected the verificationist fallacy carried over by Von Mises from deterministic theories to probabilistic ones. In Popper's view, our judgements of propensities are not inferences from observing actual frequencies, but fallible hypotheses which are subject to test and, possibly, elimination. The problem is, how do you eliminate a probabilistic hypothesis? In the deterministic case, if theory T implies an observable event E and then an inconsistent event *not-E* is observed instead, T should be rejected.[24] Not so if T is probabilistic, such as that the probability of a coin coming up heads equals 1/2. Our refuting evidence would be in the shape of a series of tosses yielding a certain actual relative frequency of heads. We have seen that even on the assumption of p(heads)=1/2, it is still more likely than not that r(heads)≠1/2 in any finite sequence. There are statistical techniques for evaluating how likely a certain relative frequency is, given the hypothesis of p(heads)=1/2. Using a result of mathematical statistics summarised in the so-called chi-squared distribution, we can use the observed r(heads) and the number of trials n, to calculate the probability that the hypothesis is true, i.e., that p(heads)=1/2. But no matter how small that probability is, it is non-zero. Therefore, no amount of evidence can properly "refute" a probabilistic hypothesis, such as

[24] Even this picture ("the hypothetico-deductive model of scientific inference") is unbearably simplistic and is now rejected by virtually all philosophers of science. It runs foul, e.g., of the "Duhem problem": should we reject the theory or the evidence that purportedly has refuted it? But my main interest here is to show that even if one were willing to grant Popper's general "falsificationism", his account of probability would not fit in with it.

our guess of a Popperian propensity. As a result, one has to select artificial thresholds of "unacceptably low" probability (most statisticians usually select 5%) which, having being breached, mean that the theory is "as good as refuted" by the evidence. I shall argue that this "cunning plan" is intellectually quite unsatisfactory.

What influences the choice of one "refutation threshold" (or, as classical statisticians[25] say, "significance level") over another? When testing a statistical hypothesis, we are in danger of either rejecting the hypothesis when it is, in fact, true ("Type I error"), or of accepting the hypothesis when it is, in fact, false ("Type II error"). Ideally, we would like to reduce the probability of both. However, given a fixed sample size, these probabilities are in inverse relation to each other.

The usual (and accurate) response is that the significance level is chosen by weighing the potential consequences of committing either type of error against each other. True, but this only shows that, contrary to Popper's contention, when statisticians accept or reject a hypothesis, their reasoning is pragmatic rather than epistemic. Accepting or rejecting a hypothesis "as true" is an epistemic act governed by its perceived truth or truth-likeness.[26] The potential cost of making a mistake should not make any difference *vis-à-vis* "acceptance as true", although it certainly matters when it comes to making practical decisions.

Suppose that, before an election to the US Congress, a candidate wants to gauge public support for President Clinton, and has the resources to canvass 1,000 people in his constituency. The survey indicates that with 90% probability, a majority of the public approves of the president. Assuming that the candidate has no principled opinion of the

[25] "Classical statistics" usually refers to the (historically predominant) school in statistics that shares the physicalist interpretation of probability. The methodology of its founder, Roland A. Fisher, was remarkably similar to Popper's.

[26] It was Popper's own belief that epistemic decisions should rest solely on the truth or truth-likeness of a hypothesis. Although his and subsequent attempts to formalise the notion of truth-likeness, or *verisimilitude*, have been rejected, that should not prevent us from viewing truth-likeness as an informal regulative idea.

President and simply wants to curry favour with the electorate, should he defend or castigate Clinton? Misjudging the public mood might cost the election. One the one hand, there is a 10% chance that despite the appearances, a majority of the public disapprove of Clinton. If the candidate then comes out pro-Clinton, he will lose. If, on the other hand, the candidate decides that the 90% probability is not sufficient evidence of a pro-president majority and comes out anti-Clinton, he runs a similar risk. What should he decide? That depends on whose votes the candidate views as marginal. If he badly needs to secure the anti-Clinton vote and views the pro-Clintonites either as granted or, on the contrary, as a lost cause, he should take a critical stance. Otherwise, he should accept the preponderance of evidence and come out pro-Clinton. There can be no suggestion that the candidate views the poll results as more or less reliable (and therefore "accepts" or "rejects" the statistical hypothesis) depending on what his base of support is. The latter is completely irrelevant to the epistemic position he will adopt.

Sometimes a practical decision is based on a hypothesis that is viewed as almost certainly false, as illustrated by Pascal. Being, for most of his life, a renowned atheist, he nevertheless advocated confession of sins on the deathbed. It costs little, he argued, and though the probability of there being a God is extremely low, the potential reward of confessing (eternal bliss) is infinitely superior to that of failing to do so (eternal damnation). Given this background, it would be absurd to interpret Pascal's deathbed confession as his sudden acceptance of Christian dogma.

A decision to accept a theory is governed by factors that go well beyond the probability that it is judged to have, hence any attempt to find a uniform 'cut-off point' for probabilistic theories is futile. *Pace* Popper, there is no satisfactory way of falsifying probability hypotheses, which in his own terminology makes them 'metaphysical', i.e. unscientific. Probability and Popper's methodology of conjectures and refutations make very uneasy bedfellows indeed.

The conclusion arrived at by many students of probability in the 1920's and later, was that whereas relative frequencies certainly give us clues as to the probabilities of outcomes,

probability itself must be something of a qualitatively different kind from frequencies, actual or virtual. This school of thought is usually referred to as the *Epistemic Interpretation* of probability, which itself falls into the Logical and Subjectivist interpretations. They will now be considered in turns.

LOGICAL INTERPRETATION

Keynes: extending logic to uncertainty

Partial entailment

As I have indicated above, the classical interpretation of the calculus of probabilities equivocated between viewing probability as a measure of objective chance, and as a measure of ignorance. While the physicalist approaches exploited the former view, the latter was further explored by John Maynard Keynes in his *Treatise on Probability*. Its main idea is the continuity between human reasoning in situations of complete certainty and those of uncertainty. Correspondingly, probabilistic reasoning was viewed by Keynes as an extension of deductive logic. The latter was represented as a special case of the former. E.g., implication was identified with conditional probability where the probability of the antecedent is one.

Let us look at what makes the epistemic interpretation of probability possible. In Kolmogorov's tradition, the arguments of probability statements are *sets*. Part of the reason why this tradition has been such a success, is that the language of set theory is ontologically neutral, i.e., can be equally well applied to different kinds of entities. In the Classical Interpretation and in Popper, it was *events*; in Von Mises, *attributes*. But for any given set, it is possible to construct a corresponding *sentence* that uniquely describes it.[27] So, we can identify the probability of a sentence with that of the set that it uniquely specifies. Hence, the probability of the coin falling heads up can be regarded as the

[27] I assume that all the indexical terms have been fixed and other linguistic ambiguities resolved, and therefore use the terms "(declarative) sentences", "statements" and "propositions" interchangeably, unless stated otherwise.

probability of the sentence 'The coin falls heads up'. Making declarative sentences the arguments of probability functions is neutral with respect to specific interpretations. Once we establish a one-to-one correspondence between "elementary" sets and sentences, we can operate with compound sentences in exactly the same way as we did with sets, taking advantage of the fact that the set-theoretic operations of union, intersection and complementation have their sentential equivalents in those of disjunction, conjunction and negation. The universal set U (which includes all the possible states of the world) corresponds to any tautology, since a tautology is consistent with any state of the world.

Furthermore, the epistemic interpretation capitalises on the observation that declarative sentences can be used to describe mental states such as beliefs. There is only one step from this to defining probability as degree of "partial" (i.e., uncertain) belief. E.g., the probability of a coin falling heads up can arguably be defined as the degree of the belief that the coin will fall heads up. Of course, belief attribution, measurable or not, require the *subject* of beliefs. For Keynes, it was "the rational agent". Any fully rational person, having access to the same information, will have exactly the same probabilities. In other words, he viewed probability as the measure of rational belief which may or may not coincide with the actual degrees of belief of fallible individuals.

What determines a unique "reasonable" measure of partial belief? The nature of such a measure becomes clear from the emphasis placed by Keynes on the continuity between full belief corresponding to logical tautologies and contradictions, and partial belief associated with contingent sentences. When are we rationally certain of a sentence A? According to Keynes, when there is a set of sentences Γ such that $\Gamma \vdash A$. In other words, certainty is based on logical entailment. Therefore, he thought, *partial* certainty must be based on *partial* entailment, or: degree of rational belief is determined by the degree of entailment. Although Keynes also allowed that numerical measures may be assigned to logical probabilities by way of comparing them to frequencies, it was the relation of partial entailment that was for him fundamental in defining probability.

The principle of indifference revisited

How can partial entailment be measured? In this regard, Keynes did not go far beyond the Classical Interpretation. To measure degree of entailment, we must divide the antecedent into a number of equiprobable cases and calculate the proportion of those cases which *fully* entail the consequent. Hence, this approach relies on the notion of equiprobability that requires the old Classical tool, the Principle of Indifference. However, aware of the paradoxes associated with the principle, Keynes attempted to exclude all those cases that give rise to them: in particular, all the continuous cases, especially the so-called geometric probabilities (victims of Bertrand's paradoxes), as well as all the countable cases where the subdivision into the "elementary" outcomes could not be made "naturally". That left the applicability of the Principle of Indifference, and therefore of the probability calculus, severely limited. Only a small sub-class of events (or sentences that describe them) could be ascribed *numerical* probability (primarily those already used as their favourite examples by the Classical writers). But even comparative probability judgements (i.e., that A is more probable than B, or vice versa, or that they are equiprobable) could be made only between directly comparable events belonging to the same "probabilistic chains". Between events from different chains, no probability judgements could be made. In technical terms, the relation "more probable than..." was, according to Keynes, only partially ordered. Hence, questions like "What is more probable, a white Christmas next year, or The Spiteful Changeling coming first in the 2:14 race tomorrow?" would be viewed by Keynes as meaningless as they ask for the comparative probability of incomparable events.

The untenability of this view is illustrated by the fact that bookies would gladly take bets on both of these events and, moreover, offer different odds for each one. This indicates that they view the probabilities of these two events as directly comparable.[28] And if the relation

[28] It may be said that what bookies are concerned about in offering odds for punters to bet on is not expressing their own subjective probability but obtaining a pool of bets that yields them a profit. While this is undoubtedly the case, in order to form an opinion on the likely profitability of the pool of bets, a bookie needs to have *some* idea of the likelihoods of the events he is offering odds on. Of course, in general the odds he offers to the punters will not reflect his own assessment of the underlying probabilities. The point is, nevertheless, that without forming *some* probability judgement on the underlying events, the bookie would not be in a position to offer *any* set of odds to the punters.

"more probable than..." *is* completely ordered, it can be proven that one can calibrate (albeit non-uniquely) the scale from 0 to 1 so as to give a less probable event a numerical measure which is smaller than that corresponding to a more probable event. In other words, if we can make *comparative* probability judgements between any two events, we can also ascribe *numerical* probabilities. A different question is whether such judgements can be made "rationally".

Carnap: uncertainty without indifference

Despite its obvious limitations, the importance of Keynes's approach is that it developed Leibniz's tradition of probability calculus as a *logic* - the logic of incomplete knowledge. From a purely logical perspective, it was taken further by Rudolf Carnap who sought to break the link between "partial entailment" and the Principle of Indifference that was the downfall of Keynes's own theory. Carnap starts (in *Logical Foundations of Probability*, 1950) by stating the need to disambiguate the concept of probability. Unlike most other writers on probability who favoured either a physicalist or an epistemic interpretation, according to Carnap, the term 'probability' has *two* legitimate uses: as a measure of objective chance, and as a measure of rational belief (or degree of confirmation). It is this latter sense that Carnap set out to "explicate".

Carnap is well aware of the shaky ground on which the Principle of Indifference, essential to the previous "explications" of epistemic probability, is founded. Most of the interesting applications are unlike the Classical urn model in that the assumption of equiprobability of elementary outcomes is unwarranted. So, for an inductive logic to be useful, it must include a way of measuring the probabilities of elementary outcomes in a more general way. He tries to do that by grounding our thinking about "elementary" outcomes in the structure of language. His analytical tool is possible-worlds semantics (cf. his 1947 *Meaning and Necessity*). Suppose we are given a formal language with recursively defined rules of formation of terms and formulae. The syntax of such a language is finite or, at most, countably infinite. The truth value of each sentence in a given possible world is a function of the references of its constituent terms in that possible world. A tautology is true in each

possible world, a contradiction – in none. All contingent sentences are true in some possible worlds and false in others. If the set of possible worlds is finite or countably infinite, than it is possible to calculate, in principle, the proportion of those worlds in which the sentence is true. That proportion gives us its probability. Thus Carnap formalised Keynes's idea of probability as the degree of partial entailment. While $B \vdash A$ (and therefore $p(B \vdash A)=1$) if B materially implies A in *each* possible world, $p(B \vdash A)=x<1$ if B materially implies A in the xth proportion of all possible worlds. That proportion gives us the "rational" degree of belief in B given A.

For a start, we must define a language relative to which a probability function is meant to represent the degree of rational belief. The building blocks of any formal language, apart from the logical connectives and quantifiers, are individual constants (names) and predicates. Names designate objects and predicates designate properties. A well-formed formula capable of a truth value has the form $A(b_1,\ldots, b_n)$, where A is an n-place predicate and $b_1,\ldots, b_n$ are names. Let us take a simple case: a language with just two one-place predicates A_1 and A_2 and one name b. Then there are two ascriptions of a predicate to a name: $A_1(b)$, $A_2(b)$, each of which may or may not hold (in the latter case, $\neg A_i(b)$ holds instead). Hence, there are four atomic sentences or their negations. Since $A_1(b)$, $A_2(b)$ are logically independent, each of the atomic sentences is consistent with both the other atomic sentence and its negation. Therefore, there are four consistent statements that fully utilise the descriptive powers of this simple language (and therefore fully describe its domain, which by stipulation is restricted to the individual denoted by b):

S_1: $A_1(b)\&A_2(b)$,

S_2: $A_1(b)\&\neg A_2(b)$,

S_3: $\neg A_1(b)\&A_2(b)$,

S_4: $\neg A_1(b)\&\neg A_2(b)$.

Since each of them completely describes a state of affairs possible in this language, Carnap called them "state descriptions".[29] Unless we are given a semantics that specifies which

[29] In general, for a two-valued logic with one-placed predicates, the number of state descriptions will equal 2^n, where n is the number of atomic sentences. In its turn, $n=km$, where k and m are the numbers

names satisfy which predicates (and therefore, which elementary sentences are true), we cannot state with certainty which of the state descriptions is the correct one. However, Carnap devises a way of calculating how *probable* each one is. First, we need a *measure* for this language. A measure in this sense is any function m that assigns a non-negative real number to the state descriptions in such a way that all the assignments add up to 1. Informally speaking, m-functions give each state of the world (as it is represented by the language) a weight. It can be an equal weight (0.25 in our example), but it need not be. Further, an m-function can be applied to *any* sentence, not just a state description. If s is a sentence, $m(s)$ is the sum of all the m-values of the state descriptions in which s holds (i.e., those where it occurs without negation). For example, $m(A_1(b))= m(S_1)+ m(S_2)$, while $m(\neg A_2(b))= m(S_2)+ m(S_4)$, etc.

Next Carnap defines the notion of a **c-function**. A c-function ("c" for "confirmation" and, sometimes, "credibility") is a two-argument one: $c(h,e)=m(h\&e)/m(e)$. A c-function is the measure of the set of state descriptions in which h holds jointly with e, as a proportion of the set of *all* state descriptions where e holds. It is clear that for every syntax, there will be infinitely many c-functions corresponding to the infinitely many ways in which different m-functions assign weights to the state descriptions. Since all c-functions satisfy the axioms of probability calculus and therefore meet the minimum requirements of intellectual consistency, they are all suitable explicants of the notions of probability as degree of confirmation. On one level, Carnap departs from Keynes by taking m-functions (which are unconditional probabilities) as primary and c-functions (which are conditional probabilities) as derivative. However, when c-functions are considered by themselves, it is conditional probability that is primary and unconditional probability that is derivative. Indeed, probability *simpliciter* is probability conditional on a tautology: $c(h)=_{df} c(h,t)$. Since a tautology is satisfied by *all* the state descriptions, unconditional probability gives us the widest possible reference class for calculating the measure of a sentence.

---

of predicates and of names.

The explication of 'probability-2' through the notion of *c*-functions was the subject of Carnap's later work *A Continuum of Inductive Methods* (1952). Since there are infinitely many ways in which an infinite set (of sentences in a language) can be mapped onto the interval [0, 1], there is an infinity of *c*-functions as well. Correspondingly, there are infinitely many ways in which the probability calculus can be used to derive inductive generalisations from particular instances.

Nevertheless, Carnap sees an advantage in giving a privileged status to a *c*-function that assigns equal weights to indistinguishable (according to a certain criterion) possibilities (a version of the Principle of Indifference). But such a "symmetrical" *c*-function is not unique, since the criterion of indistinguishability is not unique. Formally, a symmetrical *c*-function is one that is based on a symmetrical *m*-function. The latter "treats all individuals on a par" (Carnap 1950, 485). This means that all the isomorphic state descriptions (i.e., those that feature all the same predicates and differ only with respect to ascribing them to different individual constants) are given the same measure. If we expand the simple language described above to two individuals, we shall get a set of sixteen state descriptions, some of which will be isomorphic, including the following two:

$$A_1(b_1)\&\neg A_2(b_1)\&A_1(b_2)\&A_2(b_2)$$

$$A_1(b_1)\&A_2(b_1)\&A_1(b_2)\&\neg A_2(b_2).$$

Since they differ only with respect to which of the individuals is and which is not ascribed the predicate A_2, all the symmetrical *m*-functions must assign them equal values.

A disjunction of isomorphic state descriptions is dubbed by Carnap a "structure description". A structure description can be viewed as a disjunction of all those state descriptions that assign each property to a given number of individuals (with two individuals, to both, to one, or to neither).

Carnap's attempt to narrow down the class of acceptable *c*-functions still further ran into trouble. Shall the underlying *m*-functions ascribe equiprobability to structure descriptions, or to state descriptions? The former *c*-function is dubbed by Carnap c^*, and the latter – $c^\dagger$. In general, these two approaches will yield different results. In a two-predicate, two-individual constant language, there will be ten structure descriptions and sixteen state descriptions. For example,

$m^*(A_1(b_1)\&A_2(b_1)\&A_1(b_2)\&A_2(b_2)) = 1/10$, while

$m^\dagger(A_1(b_1)\&A_2(b_1)\&A_1(b_2)\&A_2(b_2)) = 1/16$,

since this sentence describes one out of ten structure descriptions but one out of sixteen state descriptions.

As we can see, Carnap's proposal falls prey to the same difficulties that render the classical Principle of Indifference unworkable – namely, the impossibility of finding a unique symmetry principle. In fact, the proposal fails on more than one level. Not only do we have a variety of symmetrical *c*-functions within a given language, but the values that each *c*-function obtains depend on the choice of a language.

Take an urn containing a certain number of balls, some of which are white. Suppose we know nothing further about the composition of the urn by colour. What is the probability of drawing a white ball? If, apart from 'white', the only elementary predicate is 'coloured' (or 'non-white'), then the probability assigned to drawing a white ball must be ½. However, if 'coloured' is further subdivided into, say, 'red', 'blue' and 'yellow', then the probability of drawing a white ball drops to ¼, since the total number of elementary predicates is now four.

In other words, even in such a seemingly simple case there in no unique way to partition the probability space into equiprobable outcomes. The same happens when we devise a formalised language to measure the probabilities of outcomes in the modal-logical

approach. The measures that result from adopting two languages with different lists of elementary predicates ("white" or "coloured", or "red", "blue" and "yellow", etc.) will be completely different. What Carnap's approach does is simply bring into the open the language-dependence and therefore arbitrariness of any formal, or "logical", attempt to define the "rational" degree of belief.

Two alternative responses are possible to this: some, as Popper, suggest that, logical probability aside, the only other legitimate sense of "probability" is the Physicalistic one. Others bite the bullet and argue that the epistemic interpretation can only be based on real, as opposed to "rational", or "warranted", degrees of belief. But while it is clear that "partial entailment" and therefore "degree of rational belief" does, indeed, follow the axioms of probability calculus (after all, in Carnap's approach it is just a relative frequency), why should "degree of belief" *simpliciter* be constrained by the same axioms?

CHAPTER 2: SUBJECTIVE BAYESIANISM

There are two ways of answering the question at the end of Chapter I. One is to analyse partial belief and show that it obeys (or must obey) the axioms of probability calculus. This could be labelled the "top-down" approach (favoured by most of the Bayesians). In the alternative "bottom-up" approach (followed by de Finetti), the probability axioms are taken as given, together with some plausible assumptions about rationality, and the behavioural patterns that are thus derived are subsequently shown to be rational. I shall examine briefly both approaches and argue that either provides a consistent justification of subjectivism.

The subjectivist approach in the interpretation of probability rests on the idea of probability being a measure of uncertainty of beliefs. Unlike the classical and logical interpretations, subjectivism denies that there exists one unique probability function that describes the "correct" degrees of uncertainty for any given proposition/state of affairs.

The logical interpretation implies a sharp distinction between the normative and the descriptive. The subjectivist, on the contrary, attempts to blur the distinction. Although, according to him, probability functions express primarily the prior *normative* constraints on human behaviour, they also, by and large, fit the *real* behaviour of individuals - so long as those individuals are being rational.[30] The last clause does not make subjectivism's claim to descriptive content empty, for according to the subjectivist, most people behave rationally (in some weak sense) most of the time, and when they do, their reasoning can be adequately described in terms of the probability calculus.

An insight into subjectivism can be given by the following observation. Suppose a coin is tossed, and a subject is asked to guess how it falls. If the guess is correct, the subject gets a fairly small but not insignificant sum of money. If the subject believes that heads is more probable than tails, he will go for heads. Why? Because that gives him, in his view, a better chance of a fixed reward. Suppose now that the subject believes the coin to be fair, that is, that heads and tails are equiprobable. Now, however, the reward is differentiated: correctly

[30] This is not to say that humans at all times unconsciously run calculations aimed at adjusting their beliefs and behaviour in accordance with probability axioms. Rather, the Bayesian would say that regardless of what 'really' happens inside, the *overt* behaviour tends to be consistent with probability calculus.

predicting tails pays more than *ditto* for heads. This time, the subject will choose tails since it gives him an even chance of winning a larger reward. Finally, let us combine the two experiments: heads is believed to be more probable, but tails yields a greater reward if guessed correctly. What will the subject choose? This question is impossible to answer *a priori*. However, it is intuitively clear that the answer depends on the relative differences between the probabilities and rewards attached to each option. By playing around with the probabilities and rewards we could create a situation where the subject is indifferent between choosing either side. While the fact of such indifference by itself does not prove the existence of a unique probability and desirability that the subject attaches to each outcome, it does show that there is an inextricable link between probability and desirability, and that individual decisions are predicated on the perceived trade-off between the two.

The above analysis hinges on the idea that a larger reward is preferred to a smaller one. If, additionally, the degrees of preference can be measured, than so can the degrees of belief embodied in a person's choice. For instance, if the reward attached to correctly guessing tails is 10 times as great as that attached to heads but the person is still indifferent between the two options, this could be explained by heads being viewed as 10 times more probable.

RAMSEY: PROBABILITY AND TRUTH

Putting your money where your mouth is

This line of reasoning was first employed, in a systematic manner, by Frank Ramsey in his seminal paper 'Truth and Probability' (1926). He starts by examining and rejecting Keynes's notion of probability as the "rational" degree of belief. According to that notion, we have direct logical intuition of the degree to which some propositions entail others. Where such intuition is lacking, the notion of probability is inapplicable. Ramsey denies that people have any such intuitions: if we did, there would be a lot more agreement than is actually observed. Further, even if we did have such intuitions, Keynes would still not be able to explain the distinction between those sets of propositions that support "rational" probability judgements and those that do not. The choice which Keynes found himself facing is this: we can either stay with the rationality requirement and severely limit the

applicability of the probability calculus, or make its applicability universal by dint of relaxing the requirement. Ramsey urges that the second option is by far the more attractive.

But if the degree of probability is not accessible through "logical intuition", how can it be ascertained? Ramsey suggests that the old dictum "put your money where your mouth is" shows the solution. The clue to the internal workings of reason is the overt behaviour of the individual. How can we measure "beliefs as bases of possible actions"? (Ramsey 1926, 34) Ramsey holds that

> The old-established way of measuring a person's belief is to propose a bet, and see what are the lowest odds which he will accept. This method I regard as fundamentally sound; but it suffers from being insufficiently general, and from being necessarily inexact. It is inexact partly because of the diminishing marginal utility of money, partly because the person may have a special eagerness or reluctance to bet, because he ether enjoys or dislikes excitement or for any other reason, e.g. to make a book (Ramsey 1926, 34).

Modern Bayesian decision theory is precisely an attempt to address these concerns in a systematic manner. However, Ramsey himself gives a good approximation to how these problems could be tackled. If money can't buy you probability, what will?

The St Petersburg Paradox

The precision with which a betting arrangement is used to measure degrees of belief will increase if as rewards we use not money per se but the units of value of money (and of other goods) to the individual. The reason why money will not do, is that money, along with any other commodity, suffers from *diminishing marginal utility*: each next unit brings smaller satisfaction than the previous one.

This phenomenon was first discussed by Nikolaus Bernoulli in the early 18th century. In his correspondence from St Petersburg (thus the appellation "St Petersburg Paradox") he described the following situation. Imagine a gamble where a fair coin is flipped repeatedly, and if heads occurs *for the first time* at the *n*th outcome, the player gets 2^n units. The expected value of this gamble is infinite.[31] It had previously been assumed that a rational

[31] On a fair coin assumption, the expected monetary value (the sum of payouts weighted by their

individual should be prepared to pay for a gamble any amount up to its expected value (i.e., pay up to £1 for a 50% chance of winning £2). Indeed, the idea that most people will only engage in a transaction if what they get in return is worth to them at least as much as what they pay back, has both intuitive appeal and vast anecdotal evidence on its side. [32]However, we would not see many people queuing up to pay all they have got for the right to participate in the gamble described above. Clearly, people tend to prefer a certain prize to a gamble of equal expected monetary value.

The apparent contradiction encapsulated in the St Petersburg paradox is easily explained if the subjective value of money is increasing at a lower rate than the quantity of it, because that decreases the weights attached to the less probable but more rewarding (in money terms) outcomes (namely, heads after a long series of tails). This is the solution proposed by Daniel Bernoulli in 1738.[33] Suppose, the subjective value of money to an individual equals the square root of its quantity. Then the St Petersburg gamble for that individual will look as follows, with *n* indicating the first occurrence of heads:

---

probabilities) is 1/2*2+1/4*4+1/8*8+..., i.e., an infinite sum of "1"s.

[32] Of course, there are risk-loving individual who are prepared to pay a premium for the thrill that taking a gamble affords them. This can be accommodated by adding the subjective value of the thrill to the subjective value of any expected monetary gain. For simplicity, we will assume that the majority of rational individuals are either risk-neutral or risk-averse, that assign either zero value or negative value to the 'risk thrill factor'.

[33] It should be noted that Daniel Bernoulli's solution is not sufficiently general, since the St Petersburg gamble can be redefined so as to be immune to it. See below for a discussion of the general conditions under which the St Petersburg paradox does not arise.

Table 3: The subjective value of a St Petersburg gamble

n	Money value of win	Subjective value of win	Probability of win	Expected subjective value of win
1	2	1.41	0.500	0.707
2	4	2.00	0.250	0.500
3	8	2.83	0.125	0.354
4	16	4.00	0.063	0.250
5	32	5.66	0.031	0.177
6	64	8.00	0.016	0.125
7	128	11.31	0.008	0.088
8	256	16.00	0.004	0.063
9	512	22.63	0.002	0.044
10	1,024	32.00	0.001	0.031

Note: Probabilities have been rounded up to the third decimal point.

Figure 3: The subjective value of a St Petersburg gamble

Note: The subjective value of having one unit of money with certainty is assumed to equal 1.

As the table and graph make clear, the expected subjective value of heads coming up on the *n*th toss converges to zero rather quickly, therefore the expected subjective value of the gamble, which is simply the sum of the values in the last column, is finite. In this particular instance, the subjective value of the gamble equals just over 2.41. In fact, this value

already obtains to the precision of two decimal points at n=20, and does not change to the 10th decimal point after n=80. Although the numbers would be different if the subjective value of money was related to its quality by a different decreasing function (Daniel Bernoulli himself used the logarithmic function), the principle remains the same.

However, as mentioned above, this solution is insufficiently general. Let us consider the *Augmented St Petersburg Gamble*: the player receives 2^{2n} (rather than 2^n, as before), if a head occurs first at the nth toss.[34] Then the expected subjective value of the gamble is $2*1/2+4*1/4+\ldots=1+1+\ldots=\infty$, as before. It is easy to see that for every strictly increasing function linking quantity of money to its subjective value, there is a modification of the St Petersburg gamble that makes its value infinite. The only general way to make the infinite value of the gamble impossible is to find a function that is *bounded above*. In such a case, even as the quantity of money tends to infinity, its subjective value cannot exceed a certain finite (although arbitrarily high) value. This was clearly not the case in Bernoulli's solution, as the logarithmic function is strictly increasing. An example of a suitable function is

$$u=m-100/n,$$

where m is a constant. It is easy to see that the value of u converges to m as n tends to infinity, thus eliminating the paradox.

Despite the insufficient generality of his own solution, Daniel Bernoulli showed that in constructing a probabilistic theory of human reasoning and behaviour, we need to calibrate the range of all possible states of affairs with respect to the *subjective* (rather than monetary) values of the goods available in them.[35] For example, in a simple world where X cares about only two goods, A and B, and assuming that there is only one indivisible unit of each, the following four propositions describe the whole range of options available to

[34] This modification was suggested to me by Colin Howson, although I have been informed that the initial reformulation along the lines described here is due to Karl Menger (who referred to it as the 'Super St Petersburg Gamble'.

[35] Of course, it is possible to avoid substituting utility for money, by stipulating that all probabilities below a certain value are in effect zero. However, this "solution" seems to me entirely unsatisfactory, as it is *ad hoc* and creates severe conceptual problems, for instance with additivity.

individual *X*: "*X* gets neither *A* nor *B*", "*X* gets *A* but not *B*", "*X* gets *B* but not *A*", and "*X* gets both *A* and *B*". If we can ascribe numerical values to the value to *X* of each of these propositions holding, we can use bets to find *X*'s degrees of belief in *all* the propositions (not only the above four).

Extracting "subjective value"

Ramsey starts the calibration with the notion of an "ethically neutral proposition" - one whose truth or falsity is indifferent to the individual:

> More precisely an atomic proposition *p* is called ethically neutral if two possible worlds differing only in regard to the truth of *p* are always of equal value; and a non-atomic proposition *p* is called ethically neutral if all its atomic truth-arguments are ethically neutral (Ramsey 1926, 38).

This notion allows us to define the probability 1/2 in an ethically neutral proposition thus. If an individual has a preference between two 'worlds'[36] *a* and *b* holding, but has no preference between the two options (1) *a* if *p* is true, *b* otherwise; and (2) *b* if *p* is true, *a* otherwise, then the individual's degree of belief in *p* is 1/2.[37]

Once the class of ethically neutral propositions believed to the degree 1/2 ("EN½") is available, we can infer the equality of utilities from the indifference between the gambles contingent on the truth value of an EN½, where the gambles involve the truth values of the propositions whose utilities we are trying to measure (following Ramsey, I shall use the same letter for a world and its utility).

If *p* is an EN½ and the options "*a* if *p* and *c* otherwise" and "*b* if *p* and *d* otherwise" are equivalent, then *ab=cd* and *a-b=c-d*. If we know the "distance" between the left hand-side options, we know it for the right hand side as well.

---

[36] According to Ramsey, a 'world' is a complete set of propositions (such that each proposition enters the set either as itself or as its negation) – in modern terms, an ultrafilter. This notion is very similar to Carnap's "state description".

[37] Hereafter I shall use the later (and now universally established) term "utility" for Ramsey's "value" - not only for consistency with further passages, but also to avoid confusion with the truth values of propositions. Also, I use notation slightly different from Ramsey's.

Subjective Bayesianism

Ramsey shows that it is sufficient to postulate the numerical utilities of just two propositions (which will give us a metric) in order to calibrate the whole range by iteratively using this procedure. And once we have defined the utilities, the probabilities follow. If an individual is indifferent between a for certain and a gamble that yields b if p (which need no longer be ethically neutral) is true and c otherwise, then his degree of belief in p equals $(a\text{-}c)/(b\text{-}c)$. This ratio represents the odds at which the individual would bet on p, and supposing that c=0 (i.e., the loser gets zero utility), it simplifies into a/b - the ratio of the price the individual is prepared to pay for the right to gamble to the potential payout.[38]

Further, Ramsey defines conditional degree of belief. Suppose an individual is indifferent between the options (1) a if q is true and b otherwise; and (2) c if both p and q are true, d if p is false and q is true, b if q is false. Then the degree of belief in p given q equals $(a\text{-}d)/(c\text{-}d)$.

Ramsey's work laid the foundation for both approaches to justifying subjective probability mentioned above. On the one hand, he sets out such a justification in the context of betting; on the other, he shows how a person's overt behaviour can be represented in terms of (subjective) probability and utility. I shall look at these alternative approaches in more detail through the work of Bruno de Finetti and Leonard J. Savage, respectively.

DE FINETTI: FROM BELIEF TO PROBABILITY

Taking a gamble

Let us recapitulate briefly Ramsey's line of reasoning. He started by accepting Keynes's premise that probability is an epistemic notion analogous to truth (the main difference being that in probabilistic inferences the confidence in the truth of the conclusion was only partial). He then pointed out the inconsistencies in Keynes's account arising from an insupportable, in his view, distinction between those epistemic states related to incomplete certainty that were logically warranted, and those that were not. His next step was to show

[38] The extra complication represented by the term c is due to the fact that there is no reason to postulate zero utility in the case of a lost bet, since the utility of an outcome of a gamble is no longer identified with the monetary payoff that obtains

that subjective degree of belief was adequately expressible and measurable, and therefore - a respectable extension of Keynes's view. The direction of reasoning is from subjective probability to degree of belief, and the argument amounts to showing that probability functions with only minimal a priori constraints on their values have natural interpretations. Savage, whose contribution is discussed in the next section, further developed Ramsey's line of reasoning.

De Finetti, on the contrary, starts by looking at degree of belief, and shows that, whenever it is coherent, it is a probability. He does that by showing that those belief systems that do not satisfy the axioms of probability are non-rational, in the sense that they result in behaviour that is, by universal standards, sub-optimal. The mechanism by which de Finetti does that later came to be known as the Dutch Book argument.

Unlike Ramsey (and the majority of Bayesians since), de Finetti believed that subjective degree of belief was a much more robust notion than that of utility, and therefore tried to explicate the former without recourse to the latter. In his view, the problems associated with the non-linearity of the subjective value of money could be overcome by choosing stakes that were neither too large nor too small, given the individual's material circumstances. With the range of stakes and payoffs chosen correctly, money bets could adequately represent the individual's degree of belief.

In outline, the argument goes as follows. Take a contingent event or state of affairs E - say, the outcome of an experiment. Further, take a statement or hypothesis a that describes E. One can interchangeably talk of E occurring or a being true. Let $cr(a)$ be an individual's degree of belief in a being true. Then, if $cr(a)$ does not obey the axioms of probability, there is a sense in which the individual is not being rational.

The argument involves two steps. The first one is to show what the correlation is between $cr(a)$ and the patterns of betting on E occurring (or, which is the same, on a being true). The second step is proving that those patterns of betting that violate probability axioms are incoherent. On the more common interpretation, introduced by de Finetti himself, the non-rationality of probability-violating patterns of betting consists in the fact that a book can be made against the individual that leads to a sure loss no matter what the outcome of the

experiment is, and that to expose oneself to such a book would clearly be non-rational. However, as we shall see, this is neither the only nor the best interpretation.

It is incontestable that there is some link between what people believe and what people are or would, under appropriate (if sometimes unattainable) conditions, be prepared to bet on. A bet on a consists in paying a sum of money (the *stake*[39]) for the right to receive a (possibly different) sum (the *payoff*), provided that a is true. If it is not, the stake is forfeited. The less a person believes in the truth of a, the higher payoff will he require in order to be willing to bet on it for any given stake (i.e., in order to offset the risk of losing the stake). In the absence of coercion or gambling addiction, the minimum condition (not to be confused with the necessary condition) for accepting a bet is that the trade-off between the net gain resulting from a being true, and the net loss resulting from it being false, be 'fair', in the sense of giving no advantage to either side. The trade-off between potential gain and potential loss is traditionally measured as their ratio and referred to as the betting odds. In other words, the odds are nothing other than the price of a gamble, in terms of net gain: for every stake s, the potential net gain is ks, where k is the odds. The following table summarises any bet with the odds k:

a	**net gain**
true	ks
false	$-s$

A set of odds is perceived as fair if the punter is indifferent between betting, with those odds, on a, or with the reverse odds, against a: $k(\neg a)=_{def}1/k(a)$. Clearly, the odds acceptable to a punter indicate his degree of confidence in the truth of a: the greater the odds demanded, the less likely is a perceived to be true. For instance, if the punter believes that a coin is unbiased, he will accept what is called 'even odds' $k=1$ (i.e., double or nothing). If the coin is perceived as biased towards heads, the punter would demand higher odds should he accept to bet on tails, but would probably accept lower odds betting on heads. Centuries of betting practice confirm a negative correlation between the perceived

[39] Often, the term 'forfeit' is used instead, whereas 'stake' is sometimes applied to the total pot of money that is at risk, namely the forfeit plus the potential net win. However, I shall use the terms as defined above.

likelihood of an event and the odds required to induce a punter to bet on it. However, is that correlation sufficiently strict?

That the odds be fair is not a necessary condition for accepting a bet (in fact, it may not even be a sufficient condition). Obviously, given a disposition to engage in gambling at all, anyone would gladly accept odds biased in his own favour - for example, a two pounds' potential net gain on a one pound stake when betting on what is perceived as an unbiased coin.

However, if the punter does not know what constitutes bias in his favour, he will have an incentive to be as 'minimal' in his demands as possible. Otherwise, if the declared fair odds were different from those that are actually perceived as fair, a greedy punter could be penalised by being required to bet the 'wrong' way. To sum up, a real or hypothetical betting arrangement where the punter does not know which way he will be required to bet, serves to elicit the odds which, in his view, are fair in the sense of conferring no advantage to either side.

The betting situation may also be described slightly differently. As a result of a successful gamble, a punter ends up with a pot of money composed of what he put on the table himself (the stake), and what he won from the other party (the net gain). This total money can be dubbed the *gross payoff*. We can now define the *price of betting* as that stake which, at the given odds, is required to obtain a given gross payoff. Let p be the price of one unit of a potential gross payoff. Then, if the target payoff is r, the total stake must be pr. The total bet can be summarised in the following table:

a	**stake**	**gross payoff**	**net gain**
true	pr	r	$r-pr$
false	pr	0	$-pr$

Let us compare the expressions for net gain in the two tables. First of all, the ratio of net gains when a is true and when it is false, equals $-k$ in the first representation and $-(1-p)/p$ in

the second. Hence, $k=(1-p)/p$. This equation shows the unique, one-to-one correlation between the odds and what is referred to as the betting quotient.

Whenever we are given a set of odds, the corresponding betting quotient is easily recoverable through the reverse transformation $p=k/(1+k)$. This is a continuous, strictly increasing transformation that maps the scale of odds onto the closed unit interval.

Quotients possess a number of valuable properties. They clearly satisfy the concern cited above: the point of indifference between a hypothesis and its negation (1/2) becomes equidistant from being certain that the hypothesis is true (1) and from being certain that it is false (0).

More generally, the quotients in betting on a hypothesis and in betting against it, sum up to 1. Remember that by definition, the odds on and against a are the inverse of each other: $k(\neg a)=1/k(a)$. Also by definition, $p(a)=k(a)/(1+k(a))$. Combining the two expressions, we get $p(\neg a)=(1/k(a))/(1+1/k(a))$. Simplifying this equation gives us $p(\neg a)=1/(k(a)+1)=1-p(a)$. But this feature of complementation with respect to 1, where the two hypotheses are contradictories, is also a property of probability functions.

As a related feature, quotients also have the same range as probability functions. Therefore, should degree of belief be represented by betting quotients, one can directly compare the former's properties with those of probability. This is where the Dutch Book Theorem (or the 'Ramsey-de Finetti' Theorem) can be invoked.

A Dutch book is such a betting arrangement (as described above) that the punter loses with certainty, regardless of the outcome of the truth of a. It is fairly easy to prove that a Dutch book can be constructed if and only if betting quotients (and therefore degrees of belief) are not probabilities.

On the more popular 'standard' interpretation, the proof amounts to showing that violating any of the four probability axioms makes the punter susceptible to a Dutch book and therefore - to a sure loss, which cannot be rational. The expositions of the Ramsey-de

Finetti Theorem that are based on this interpretation, assume that it is the bookie who chooses which way the punter is required to bet, and who exercises that choice to the punter's disadvantage. The punch-line then is that the punter would be irrational to expose himself in such a fashion. This interpretation contains a number of behavioural assumptions that Howson and Urbach, among others, find either questionable or unnecessary.

Therefore, Howson and Urbach adopt an alternative way of couching the proof. They argue that it is immaterial which way the bet goes; the important thing is that, if the punter's betting quotients do not follow the axioms of probability, one of the sides will get a certain advantage. The contention now is not that the punter's behaviour is 'economically' irrational, but that his definition of fair odds is logically inconsistent. Indeed, if we assume that a betting quotient represents subjectively fair odds, then it must confer zero net advantage to either side. If a positive net gain to one of the sides does result with certainty from a set of odds, then they cannot have been fair odds, contrary to the assumption. I shall illustrate the point by proving the Additivity Axiom. The proof follows that in Howson and Urbach (1993, 80).

Take two separate bets on two mutually exclusive hypotheses, a and on b, designating the betting quotients p and q, respectively. Let us select the unit of currency so that the value of the gross payoff is 1 in each bet. Then, given the definitions above, the following net gain matrix obtains:

a	***b***	**net gain**
true	**false**	$1-p-q = 1-(p+q)$
false	**true**	$1-q-p = 1-(p+q)$
false	**false**	$-q-p = -(p+q)$

The first two cases correspond to the conditions of truth for the disjunction of a and b (remembering that the two hypotheses are mutually exclusive, which rules out the 'true-true' case), and the third case represents the falsity condition for the disjunction. Therefore,

the matrix can be re-written as follows:

avb	**net gain**
true	1-(p+q)
false	-(p+q)

Suppose now that the betting quotient on *avb* (call it r) is different from the sum of the betting quotients on *a* and on *b*. In other words, assume that $r \neq p+q$. As has been shown above, the quotient in betting against the disjunction is 1-r. Then the net gain matrix for the joint bet on *a*, *b*, and against *avb* takes the following form:

avb	**net gain**
true	1-(p+q) - (1-r) = r-(p+q)
false	-(p+q) + 1-(1-r) = r-(p+q)

The net gain does not depend on the truth of *a* and *b*. If $r>(p+q)$, the punter is assured a profit, come what may. If $r<(p+q)$, he is assured a loss, come what may. In either case, the betting quotients involved cannot have been based on fair odds, because those, by definition, are designed not to give certain advantage to either side.

It easy to show, by a virtually identical argument, that fair betting quotients are also countably additive. It is slightly harder to show that fair betting quotients satisfy the last axiom: Conditional Probability. First, we must introduce the notion of a conditional bet. A bet on *a* conditional on *b* is such a bet on *a* that it is allowed to proceed only if *b* is true. If *b* is not true, the bet on *a* is off and all forfeits are returned. As an example, take the bet on Japan reducing its interest rates in April, conditional on the US doing likewise in March. Another example is the bet on Arsenal FC winning the Premiership this year, conditional on it winning its next game. If the US does not reduce its interest rates in March, or if Arsenal fails to win its next game, the main bet is off and the forfeits are returned. Since there is no principled bar preventing a punter from entering into such a bet, there must be a

matter of fact about what odds (and therefore what quotient) he would consider fair for that bet. Let us denote that subjectively fair quotient p, and denote the arbitrary stake s. Then the net gain matrix, given the definition of a conditional bet, takes the form of

a	***b***	**net gain**
true	true	$s(1-p)$
false	true	$-sp$
true or false	false	0

Next, let us look at the simultaneous non-conditional bets on *a&b* and against *b*. Let q be the betting quotient on *a&b*, and let r be the betting quotient on *b*. Suppose the stake chosen on *a&b* is $r>0$, and the stake chosen on *b* is q. Then the net gain matrix for the joint bet on *a&b* and against *b* is as follows (given the truth conditions for the conjunction, we can disregard the combination of *a&b* being true and *b* being false):

a&b	***b***	**net gain**
true	**true**	$r(1-q) - q(1-r) = r(1-q/r)$
false	**true**	$-rq - q(1-r) = -q = -r(q/r)$
false	**false**	$-rq + qr = 0$

Comparing the two tables above, it becomes clear that they are identical for $p=q/r$ and $s=r$. In other words, to each pair of non-conditional bets as above, there corresponds a conditional bet with the quotient determined by the former of these two equalities. The latter equality is not restrictive, since s was assumed to be arbitrary. As Howson and Urbach point out,

> Hence, if you were to state a fair conditional betting-quotient which differed from q/r, you would implicitly be assigning different conditional betting-quotients to the same hypothesis (Howson and Urbach 1993, 83).

Indeed, by betting simultaneously on *a&b*, against *b*, and against *a* conditional on *b* (where the forfeits are as above), we would get the following net gain matrix:

a&b	***b***	**net gain**
true	**true**	$r(1-q/r) + r(p-1) = -q+rp$
false	**true**	$-q + rp$
false	**false**	$0 + 0 = 0$

We would expect two opposite bets with the same stake to cancel each other out in each possible case, whether or not one of them is made implicitly. But the only way in which it happens here, i.e., the only way in which net gain is zero regardless of the truth values of *a* and *b*, is by *q* being equal to *p*. This, obviously, entails $p=q/r$, where $r>0$.

Howson and Urbach believe that it is insufficient to show that violating $p=q/r$ leads to a non-zero net gain in the triple bet described above. Their argument is that in this arrangement, violating the equality does not yet doom one of the parties to a certain negative net gain. Indeed, the net gain for both parties is zero in the case where *b* is false. They propose to close this loophole by adding a further bet: on *b* with stake $q-pr$. However, in my view, this is unnecessary. Demonstrating that a bet on and against the same proposition, with identical forfeits, gives a non-zero net gain in at least some truth-value assignments, is enough of a reduction *ad absurdum* to reject the condition that yields such a result.

Thus, we have seen that a system of partial beliefs is consistent only if the quotients by which they are measured are formally probabilities. But having a probability metric is a sufficient, as well as a necessary, condition for a system of beliefs to be consistent. Indeed, when partial beliefs are measured by betting quotients which, in turn, obey the probability calculus, no betting strategy leads to a guaranteed net loss.

From belief to probability: are we there yet?

The betting approach does not convince everybody, however. Why should we assume that the machinery involved in betting be non-distortive with respect to beliefs? People might

be attracted to the betting process itself, and thus willing to accept bets at what they believe to be, strictly speaking, unfair terms. Similarly to the problem of measurement in quantum physics, where the physical characteristics of a system are affected by the very act of measurement, betting under the rather artificial conditions described above, might distort the beliefs being measured.

This charge can be met in two ways. Firstly, we are concerned not so much with the actual values of subjectively fair betting quotients, but with the structural relationships between them - to wit, with whether or not they satisfy the axioms of probability. Even if a punter is attracted to betting to the extent of being willing to incur a loss, he will still try to minimise the average loss per bet, and that is enough to justify the axioms of probability.

Secondly, the objection does not hold for hypothetical bets, where the 'bettor' has no incentive to depart from what he honestly believes to be the fair odds. While the elicitation of values becomes more problematic than with actual bets, the problem of structural distortion does not arise.

A more specific concern is with the technique of aggregation of bets which makes a Dutch book possible. The threat of a Dutch book arises because the bookie may be able to bundle together a set of bets severally acceptable to the punter, and concoct a joint bet which with certainty leads to the punter's net loss. But, it can be argued, joint acceptability of severally acceptable bets is highly dubious. Patrick Maher provides the following example.

Suppose that, late at night, you are stranded far from home with only 60 cents in your pocket, while the bus fare home is $1. If you are then offered to bet, with the quotient of 0.6, on a coin that you believe to be fair landing heads, it would be rational of you to stake your 60 cents. Indeed, you have a 50% chance of winning $1, which would enable you to take the bus home. If you accept the bet, there is a 50% probability of having to walk home, while if you refuse, that probability is 100%.

> Since you believe the coin to be fair, you would be just as eager to bet your 60 cents on tails. By the aggregation principle employed in the Dutch Book argument, you would then be willing to accept a joint bet on heads and tails, each at the same 0.6 quotient. But that would mean a sure loss for no

> good reason. In being willing to bet at less than even money on either heads or tails, you are merely being sensible; but you would certainly have taken leave of your senses if you were willing to accept both bets together. Accepting both bets, like accepting neither, means you will have to walk home (Maher 1993, 96).

Maher than considers a number of ways in which a proponent of the Dutch Book argument might respond. The most obvious is to say that in the putative counter-example aggregation fails to work only because the sums of money involved are not the main consideration. Getting home easily and safely is more important than avoiding being ripped off by the bookie. The utility of a 50% chance of a bus ride outweighs the disutility of an unfair bet, whereas there is no such mitigation in the case of a joint bet that leaves no money for the bus home, anyway. It was precisely in order to filter out such extraneous influences that the Dutch Book theorist was so careful in devising his (admittedly artificial) set-up.

Maher's rejoinder is that in assuming the existence of utilities (the utility of money versus the utility of a bus ride), the Dutch Book theorist begs the question. For once we have made enough assumptions to justify the introduction of utilities, we can prove that subjective degree of belief is a probability, without recourse to the Dutch Book argument. Probabilities and utilities emerge together in a Representation Theorem. Hence, the Dutch Book argument is either unsound (for without utilities the assumption of aggregation is false), or superfluous (for when introducing utilities to justify aggregation, subjective probabilities emerge as a by-product).

Two responses are appropriate here. Firstly, as we shall see in the next section, the Representation Theorem is needed to introduce numerical utility functions, rather than the notion of utility as such. On the contrary, the notion of utility involved in the Dutch Book argument, is a qualitative one. It is used there in two ways: (1) to assert that for a certain range of forfeits and payoffs, the subjective enjoyment of money grows more or less linearly in proportion to its amount; and (2) to stipulate that in a carefully controlled set of conditions, the money involved in the betting is the bettor's only concern. These are much weaker assumptions than what follows from the Representation Theorem, and arguably can be justified to a sufficient degree by reference to a 'naive', qualitative idea of utility as subjective usefulness.

Secondly, the very counter-example that Maher thinks so decisive, is in fact spurious. If you had 60 cents but needed $1, surely you could not possibly conclude two bets with a 60 cent stake each? Hence, the counter-example to aggregation would not arise. If you were asked to bet on both heads and tails with the halved forfeits of 30 cents each, you would indeed be irrational to accept such a bet, but then if would be equally irrational to accept separate bets on either heads or tails with a 30 cent stake, since that would give you no chance of ending up with the $1 required to pay the fare.

I shall argue that, rather than (as Maher believes) supplanting the Dutch Book argument, the Representation argument supplements it. The assumptions of either approach can be questioned, but the fact that they independently lead to the same results, is quite impressive.

SAVAGE: AN ONTOLOGICAL GROUNDING

Preference

Savage also explicitly grounds probability in the behaviour of rational individuals in an uncertain world. On the world side, his basic notions are 'the **world**', 'a **state** (of the world)', and 'the **true state** (of the world)'. The meanings of these terms are 'in reasonable harmony with the usages of statistics and ordinary discourse' (Savage 1954, 9). However, Savage's ontology is "minimal" in the following sense: only those aspects of the world are explicitly included about which the individual is not wholly indifferent. Take, for instance, a world consisting only of a chess board, two players (wearing clothes), and a few candy wrappers lying around. If the players are concerned only with the game and its outcome, we can safely omit the wrappers and the clothes, from the ontology. Moreover, even in this "coarser" world there will be states between which the players are indifferent, such as the otherwise identical states in which the white bishops initially at C1 and F1 are swapped around. These pairs of states can be merged, thereby reducing the total number of states of the world by half. Such a reduction will buy us greater tractability at the cost of smaller descriptive power. This procedure can be repeated until the identity of each piece is fully exhausted by its location and function[40]. In the coarsest adequate world[41], by definition, all the states are mutually exclusive and jointly exhaustive.

We may want to "lump together" certain states that are distinguishable only in certain relevant respects. For instance, the numerous positions on the board in which the black king is under check against which it has no defence, can be gathered in a set labelled "check mate to the black". Savage calls sets of states **events**. This term is also rooted in ordinary-language use: it is customary to refer to entities like "check mate to the black" as 'events'.

---

[40] For simplicity, let us ignore the few history-dependent aspects of a chess game, such as the impossibility of rooking after the figures involved have been moved.

[41] I.e., the world where all the states indistinguishable from each other in all relevant respects (coarseness) but none of the states differing from each other in at least one relevant respect (adequacy) have been merged. Incidentally, the terms are not Savage's, nor is the chess example.

Since there are no restrictions on the sets of states that qualify as events, they range from the set that includes all the possible states (the **universal** event *S*) to the set that includes no sets at all (the **vacuous** event). Of an event that includes the true state as its element, Savage says that it "holds". The universal event holds always and the vacuous one - never. Events, just as any sets, satisfy the Boolean algebra. Following Savage, let us denote the states of the world (i.e., the elements of *S*) *s*, *s'*, ... , and their sets (i.e., the subsets of *S*, or events) *A*, *B*, *C*, ...

Those events that come about as a result of the actions of individuals *and* matter to the individuals, can be dubbed the "consequences" of those actions. Often, the consequence of one act is also the antecedent condition of another. In fact, in the world of a chess game each position is the result of a previous move (a consequence) and also the antecedent of the next move - apart from the initial position that does not result from any act, and the final position that does not form the basis of any further act (accept for the act of surrendering or agreeing a draw). However, for the sake of generality it is better not to assume that the set of consequences is a sub-set of the set of states of the world[42], and denote the set of consequences '*F*'. The following example of Savage's will be useful in clarifying this distinction.

Suppose we are making an omelette. Five eggs are already broken and in a bowl. The sixth and last egg can be either broken directly into the bowl; or broken into a saucer and then transferred into the bowl if fresh; or thrown away. Savage summarises the relation between these alternative acts, the antecedent states, and the possible consequences, in the following table:[43]

[42] Normally, we want to separate the two sets in those cases where causal relations from consequences to (future) states of the world either do not exist or can be ignored, as in one-step games. Savage's example below is a case in point.

[43] The example is slightly modified, as originally the consequence of throwing away the sixth egg when it was in fact good, also included the unnecessary destruction of one good egg.

Act	State	
	Good	Rotten
Break into bowl	six-egg omelette	no omelette, and five good eggs destroyed
Break into saucer	six-egg omelette, and a saucer to wash	five-egg omelette, and a saucer to wash
Throw away	five-egg omelette	five-egg omelette

Each consequence is uniquely determined by an act together with the state of the world that obtains. (Conversely, given the state of the world, each act uniquely determines the set of consequences.[44]) Therefore, an act can be understood as a *function* from states of the world to consequences. So, if the act of breaking the egg directly into the bowl is denoted '**f**', then

f(good) = six-egg omelette;

f(bad) = no omelette, and five good eggs destroyed (Savage 1954, 14-15).

Note that in specifying the consequences, we are concerned only with that in the description of the corresponding state that matter to the individual. For instance, should we throw away the sixth egg, we end up with a five-egg omelette regardless of whether the sixth egg was good or rotten. Whereas the state of the world that results from this act will be different depending on the antecedent state of the world that obtains (a good egg lying in the bin, as opposed to a rotten egg lying in the bin), this distinction does not enter into the description of the consequences. The act of throwing the last egg away also exemplifies

[44] Where it is seemingly not the case, we are not partitioning the acts finely enough. For instance, "break the sixth egg" is not a bona fide act: it has to be split into "break into the bowl directly" and "break into a saucer first". Analogously, two acts that have exactly the same consequences given the same antecedent states of the world, are not genuinely different. If all the relevant consequences are listed in the table above, "throw the egg into the bin" and "throw the egg out of the window" should be viewed as one act.

an important sub-species of acts - *constant acts* (i.e., those whose consequences are independent of the states of the world).

Following Savage, let us reserve the letters **f, g, h,** ... for acts, and letters *f, g, h,* ... for consequences. Then acts are arbitrary functions from the set of states *S* to the set of consequences *F*.

Human beings are constantly engaged in *choosing* between acts. To explain this observation folk-psychologically, they *prefer* to act in one way as opposed to any other. Savage reflects this explanation in postulating the relation "is not preferred to" (usually dubbed in decision theory "weak preference"). He denotes this relation "$\preccurlyeq$" and postulates that it is a *simple ordering relation* between acts. Expressed formally,

P1 For any **f, g** and **h,**

1. Either $\mathbf{f} \preccurlyeq \mathbf{g}$ or $\mathbf{g} \preccurlyeq \mathbf{f}$.

2. If $\mathbf{f} \preccurlyeq \mathbf{g}$ and $\mathbf{g} \preccurlyeq \mathbf{h}$, then $\mathbf{f} \preccurlyeq \mathbf{h}$ (transitivity).

The relations of *strong preference* and *indifference* between acts are derived in the following way:

$\mathbf{f} \prec \mathbf{g}$ iff $\mathbf{f} \preccurlyeq \mathbf{g}$ but not $\mathbf{g} \preccurlyeq \mathbf{f}$;

$\mathbf{f} = \mathbf{g}$ iff $\mathbf{f} \preccurlyeq \mathbf{g}$ and $\mathbf{g} \preccurlyeq \mathbf{f}$.

To go any further, we need the notion of preference between *consequences*. One consequence is preferred to another **iff** they are, in each state, the consequences of two acts one of which is preferred to the other. Or, formally,

$g \preccurlyeq g'$ iff $\mathbf{f} \preccurlyeq \mathbf{f}'$ when $\mathbf{f}(s)=g$, $\mathbf{f}'(s)=g'$ for every s in S.

Often we need to compare acts which have the same consequences in some states of the world and different ones in others. If *B* is the set of states in which **f** and **g** have the same consequences, these two acts are said to *agree on B*. Analogously defined is "preference on *B*" (which can also be dubbed "conditional preference"). In an extreme case, agreement on *B* may hold for any pair **f** and **g**. Let us imagine that the list of the states of the world in the omelette example is incomplete. For instance, in addition to "good egg" and "rotten egg" we also had "a time bomb disguised as an egg and about to go off". If this latter state of the world obtained, all the possible acts would have the same consequence: "no omelette, no dishes to wash". In the case of most people, this state would be ignored in decision making since in relation to it, we would be indifferent between *all* the available acts. We would be acting as if that state were impossible. Indeed, we might view it as impossible or "virtually impossible". In Savage's terminology, it is "null". Although the vacuous event (containing no possible states of the world) is null, the latter notion is wider. A **null event** is such that in relation to it, the individual is indifferent between *any* two acts.

The notion of conditional preference allows Savage to state an important informal idea - the **Sure-Thing Principle**:

> If the person would not prefer **f** to **g**, either knowing that the event *B* obtained, or knowing that the event ~*B* obtained, then he does not prefer **f** to **g** (Savage 1964, 21).

Savage illustrates the intuitive appeal of this principle with the following example. A businessman is considering buying a piece of property. He believes the outcome of the upcoming presidential election relevant to the appeal of the project. Upon reflection, he decides that he would buy the property if the Republican candidate wins, and *ditto* for the Democratic candidate. Since his decision is the same no matter who wins, he should buy the property regardless of the outcome of the election. As Savage points out himself,

> It is all too seldom that a decision can be arrived at on the basis of the principle used by this businessman, but, except possibly for the assumption of simple ordering, I know of no other extralogical principle governing decisions that finds such ready acceptance (Savage 1954, 21).

The Sure-Thing Principle derives unconditional preference between acts from preference conditional on complementary states. The intuition behind it can be extended to cover

cases where an act agrees with one act on a certain state and with a different act on the complementary state.

Suppose that an individual is going for a walk but is unsure whether to prepare for rain or for sunshine. All he cares about is protecting himself from the rain and the glare of the sun. Suppose further that at most one of the following could be taken along: an umbrella, a raincoat, or a pair of sunglasses. Which of these objects, if any, should he take for the walk? All the states of nature, available acts and their possible consequences are summarised in the following table:

Act	State	
	Sunshine	Rain
take umbrella	Protected from glare	protected from rain
take raincoat	Exposed to glare	protected from rain
take sunglasses	Protected from glare	exposed to rain
take nothing	Exposed to glare	exposed to rain

In case of sunshine, taking an umbrella is indistinguishable from taking sunglasses, and taking a raincoat - from taking nothing. In case of rain, taking an umbrella is indistinguishable from taking a raincoat, and taking sunglasses - from taking nothing. Since it gives the individual protection in all the eventualities, taking an umbrella is preferred to taking a raincoat (whose protection is limited). Analogously, since it gives some protection, taking sunglasses is preferred to taking nothing (which provides no protection at all).

Savage would argue that, in fact, the last statement follows from the preceding ones. To formalise this intuition, he postulated

P2: If **f**, **g**, and **f'**, **g'** are such that

1. in $\sim B$, **f** agrees with **g**, and **f'** agrees with **g'**,

2. in B, **f** agrees with **f'**, and **g** agrees with **g'**,

3. $\mathbf{f} \preccurlyeq \mathbf{g}$;

then $\mathbf{f}' \preccurlyeq \mathbf{g}'$.

In a similar manner, Savage uses what he views as uncontroversial intuitions about preferences between acts, events and consequences to introduce a further five decision-theory postulates. The notion of conditional preference between events is formalised in

P3: Given two constant acts and a non-null event B, one is preferred to the other given B **iff** its consequence is preferred to that of the other act.

Together with the first two postulates, this implies a stronger and more obviously useful result. If we partition an event and in each element of the partition the consequence of one act is preferred to that of another, then the former act is preferred to the latter, given the event. Moreover, if the preference between the consequences is strict in at least one of the non-null elements of the partition, the preference between the acts is strict.

Let us recall the observation that if a person is promised a prize on the condition that he guesses the outcome of a trial correctly, he will choose that outcome which he deems more probable. In this behavioural pattern, we have a bridge between preference and subjective probability. But does that pattern depend on the prize? If it does, the preference could be reversed when a more attractive or a less attractive prize is offered, thus preventing us from constructing the person's probability function. Intuitively, this seems implausible. Therefore, Savage proposes a further postulate:

P4 if an individual prefers a prize conditional on one event to the *same* prize conditional on another event, his preference is independent of the prize.

This invariance allows us to define, slightly rephrasing Savage, **Qualitative probability**:

if the identical prize is offered conditionally on events A (act $\mathbf{f}_A$) or on B (act $\mathbf{f}_B$),

and the individual (weakly) prefers that prize conditionally on B,
then A is **not more probable** than B ($A \preceq B$).

In order to be able to make stronger statements, we need to ensure that the notion of strong preference introduced before derivatively is non-vacuous. **Postulate 5** takes care of that by stating that

P5: There is at least one pair of consequences one of which is strongly preferred to the other.

Probability

Now Savage is ready to define the notion of **qualitative probability** as any relation $\preceq$ between events that, for any events B, C, D, satisfies the following conditions:

1. $\preceq$ is a simple ordering relation;
2. $B \preceq C$ **iff** $B \cup D \preceq C \cup D$, provided D is disjoint from B and C;
3. $0 \preceq B$, $0 <\cdot S$.

Postulates 1-5 jointly entail that the relation 'is not more probable than' is qualitative probability. A number of consequences that conform to the intuitive concept of probability are derivable from this result. However, a more interesting question is how this qualitative notion could be linked to the quantitative probability as developed e.g. by Kolmogorov. We want to find such probability measures that "agree" with '$\preceq$', in the following sense:

If S carries a probability measure P and a qualitative probability $\preceq$ such that, for every B, C, $P(B)<P(C)$, if and only if $B<\cdot C$; then P **(strictly) agrees** with $\preceq$ If $B<\cdot C$ implies $P(B) \not\succeq P(C)$, then P **almost agrees** with $\preceq$ The "almost" part reflects the possibility that

> $P(B)=P(C)$, though $B<C$, so that knowledge of P may imply only imperfect knowledge of $\preceq$ (Savage 1954, 34).

What are the existence conditions for a probability measure on events agreeing with qualitative probability? The main one is that the partition of the universal event S into states be almost uniform:

> An *n-fold almost uniform partition* of B is an n-fold partition of B such that the union of no r elements of the partition is more probable than that of any $r+1$ elements (Savage 1954, 34).

Among the results provable from **P1-P5**, is that

> If there exist n-fold almost uniform partitions of S for arbitrarily large values of n, then there is one and only one probability measure P that almost agrees with $\preceq$ (Savage 1954, 34).

This means that under fairly unrestrictive assumptions about the world, there is a unique numerical representation of a person's relative probability judgements. These assumptions are articulated in **Postulate 6'**:

> If $B <\cdot C$, there exists a partition of S the union of each element of which with B is less probable than C (Savage 1954, 38)[45].

In other words, the world is viewed as composed of so highly individuated states that the addition of one of them to an event does not change its place in the ordering of events according to their qualitative probability.

Qualitative probability on *events* was originally derived by Savage from preference between *acts*. Therefore, it is not surprising that **P6'** has a direct correlate for acts, **Postulate 6** (**P6'** is its logical consequence and therefore weaker):

---

[45] In other words, the relation $\preceq$ is both "fine" and "tight". It is fine **iff** for every $B>0$, there exists a partition of S no element of which is as probable as B.

Two events, B and C, are "almost equivalent", **iff** for all non-null events G and H that are disjoint from B and C respectively, $B \cup G \preceq C$ and $C \cup H \preceq B$ (i.e., the addition of a disjoint non-null event makes one of them at least as probable as the other already is). Finally, $\preceq$ is tight **iff** every pair of almost equivalent events are strictly equivalent.

> If $\mathbf{g} \preccurlyeq \mathbf{h}$, and f is any consequence; then there exists a partition of S such that, if **g** or **h** is so modified on any one element of the partition as to take the value f at every s there, other values being undisturbed; then the modified **g** remains less than **h**, or **g** remains less than the modified **h**, as the case may require (Savage 1954, 39-40).

That is, we can choose a partition of the world fine enough for an individual to maintain a strict preference between any two acts even when the "better" act is arbitrarily worsened, or the "worse" act arbitrarily improved, in any one state of the world. The preference will remain unchanged because *any one* act in a sufficiently fine partition is "insignificant".

Both in mathematical probability theory and in applications, the notion of **conditional probability** is all-important. Does personal probability have the required properties? Just as it was with unconditional probability, the question consists of two parts. Firstly, can conditional probability be defined in a way consistent with **P1-P6**? Secondly, if it can, does it have a unique quantitative representation? The answer to both questions turns out to be positive. Remember the definition of qualitative probability: $B \preceq C$ **iff** $\mathbf{f}_B \preccurlyeq \mathbf{f}_C$, where $\mathbf{f}_B$ and $\mathbf{f}_C$ are defined as choosing between identical prizes conditional on events B and C respectively. Analogously, Savage defines $B <\cdot C$ given D as $\mathbf{f}_B \prec \mathbf{f}_C$ given D. The latter relation is nothing other than a preference relation between choosing a prize conditionally on $B \cap D$ and choosing it conditionally on $C \cap D$ (provided D is not null). Since, according to **P4**, the definition of $B \succcurlyeq C$ is independent of the nature of the prize, we can dispense with acts altogether and define conditional probability purely in terms of events:

> If $\preceq$ is a qualitative probability, and $0 <\cdot D$; then $B \preceq C$ **given** D, if and only if $B \cap D \preceq C \cap D$ (Savage 1954, 44).

He then proves that $\preceq$ given D has the properties of being a qualitative probability, being fine, and being tight, **iff** $\preceq$ has the respective properties. Hence, if $\preceq$ is fine, then for any non-null D there exists a unique (conditional) probability measure $P(B|D)$ that almost agrees with $\preceq$ Moreover, it can be proven that $P(B \cap D)/P(D)$ is *also* such a measure. Therefore, due to uniqueness, $P(B|D) = P(B \cap D)/P(D)$.

In other words, under certain, fairly modest, assumptions personal probability (including conditional probability) has a unique probability measure representation. This vindicates interpreting probability as a (subjective) degree of belief and applying to the latter all the formal results of probability calculus, including Bayes Theorem.

Utility

Preference between acts hinges on two things: preference between their consequence in various states, and the perceived likelihoods of those states. The latter aspect has been arithmeticised with the help of quantitative probability. If the former could also be subjected to a similar arithmetisation, comparing different act would become a formally decidable procedure. Ramsey had already shown how numbers could be assigned to acts in a way that represents their value to individuals, thus anticipating the formal notion of utility developed by John von Neuman and Oscar Morgenstern in their *Theory of Games and Economic Behaviour*, and Savage uses their results with minor modifications.

A **utility function** $U(f)$ is a function that attaches a real number to each consequence in such a way that $U(f) \trianglelefteq U(g)$ **iff** $f \preccurlyeq g$. Defining utility for an act is trivial for constant acts (i.e., those that have the same consequence in all states): $U(\mathbf{f})=_{def}U(f)$, where f is the consequence of **f**. For non-constant acts (with a finite number of consequences), the utility of an act is the **expected utility** of its consequences, or the sum of the utilities of the consequences weighted by their respective probabilities. As a result, we can now define utility as any real-valued functions over acts such that

$$U(\mathbf{f}) \trianglelefteq U(\mathbf{g}) \textbf{ iff } E(U(\mathbf{f})) \trianglelefteq E(U(\mathbf{g})).$$

What remains to be examined is the conditions under which such functions exist. Savage shows that the existence of a utility function for any given set of acts, already proven by von Neuman and Morgenstern, can also be deduced from **P1-P6**. The uniqueness result, however, is weaker than that for probability: if U is a utility function, then so is any $U'=aU+b$, where a and b are real and $a>0$. The reverse can also be proven true: if U and U' are utilities, then there exist a and b that are real and $a>0$, such that $U'=aU+b$. In other

words, utility is unique only "up to a positive affine transformation".[46] From the point of view of decision theory and game theory (that is, where the notion of utility is primarily used), this is sufficient. In decision theory, a rational individual chooses that course of action which maximises his expected utility, and it is a mathematical fact that two functions that are positive affine transformations of each other have their maxima at the same point, so the difference in the specific numbers attached to expected utility at the optimum point turns out not to matter.

Where unique utility values would have been useful, is in inter-personal comparisons. Such areas of study as social philosophy and general equilibrium theory in economics, are concerned with which of the many productively efficient allocations of resources are socially optimal, from the point of view of "fairness" or "equity". Unless we have a unique utility function that describes *everybody's* preferences on the same scale, there is no "scientific" way of knowing under what conditions taking from Peter and giving to Paul is socially optimal.[47] However, in game theory, where an agent ("player") has to make decisions based on the anticipated strategies of the other players, the normative issues do not arise and therefore inter-personal utility comparisons are unnecessary.

To sum up, in all purely descriptive contexts the apparatus of utility is sufficient. This corroborates the Bayesian contention that analysing human behaviour in terms of utility and its correlate - subjective probability, is a well-grounded and promising research programme.

---

[46] Nevertheless, this is a stronger result than one that would obtain if each act were constant, i.e., in a world without uncertainty. In such a world, utility would be unique only up to a *strictly increasing* transformation (even non-linear, such as logarithmic, quadratic, etc.). In the microeconomic models that assume a world of complete certainty, utility is "ordinal" (i.e., merely preserves the order of preference) rather than "cardinal" (i.e., where the actual numbers have significance).

[47] Even if interpersonal utility comparisons were possible, we would still face the problem of defining "social optimality". At one extreme, it might consist in maximising the sum of all individual utilities; at the other, in absolute equality, with infinitely many intermediate solutions. As a substitute of distributive efficiency *simpliciter*, social philosophers and economists use the much weaker notion of *Pareto Efficiency*. An allocation is Pareto efficient **iff** the only way to make someone better off, is by making someone else worse off. That this notion is not sufficiently strong, follows from its almost complete neutrality with respect to any possible system of personal taxation. Virtually any change to personal tax is Pareto inefficient, even in cases where there is a clear social benefit.

KINEMATICS OF BELIEF

The argument so far has been that, in so far as it is rational, an individual's belief system at any given time obeys the probability calculus. However, the Bayesian idea is much more ambitious: namely, that rational belief change is also guided by the laws of probability. In other words, the thesis is that belief change is based on the process known as Bayesian conditionalisation.

According to Bayes Theorem, $p(h|e)=p(h)*p(e|h)/p(e)$, where h and e are any two propositions. On the subjectivist interpretation, Bayes Theorem describes the relationship between the degree of belief in a proposition conditionally on another proposition, and unconditionally. Imagine, after a long illness, regaining consciousness on a grey, rainy, moderately cold day in London. You cannot see either trees or passers-by from the window. You wonder what time of year it is (let us assume for simplicity that there are only two seasons, summer and winter). You say to yourself: 'It looks very much like winter; I would give it 90% probability. However, if I heard thunder, I would think it 60% probable that it's summer.' One can say that, since thunder is more likely in summer than in winter, it would constitute support for the belief that it is currently summer. The exact numerical degree of support, measured either as the ratio of the conditional and unconditional degrees of belief, or as their difference, is given by Bayes Theorem. In this case, a stroke of thunder would give the 'summer' hypothesis 50% support (using the latter measure).

There are two ways of interpreting this situation. One is to say, counter-factually, that had thunder been heard in your present state of mind, you *would* have been 60% confident that it's summer. The second interpretation is that, if you learn or hear nothing else of relevance in the near future, the sound of thunder *will* make you 60% confident that it's summer. On the second interpretation, we are able to draw definite predictions about people's future states of mind if we know their present states and the information that they will be exposed to.

The 'counter-factual' interpretation is a simple restatement of Bayes Theorem and, as such, is uncontroversial. Unfortunately, it is also of little practical use. Both in practical

decision-making and in scientific theorising, the question is how to change one's state of mind over time, as new information becomes available. The 'factual' interpretation, on the other hand, crucially depends on the following assumption:

> *The degree of belief that will be rational after new information becomes available, is that degree of belief which would have obtained had the same information, and no other, been available now.*

A person's probability function $p(h)$ takes as its arguments his current beliefs. We can relativise 'current' by indexing p with time, and compare the probability assignments over time. Then the assumption (usually referred to as the *Principle of Conditionalisation*) behind the 'factual' interpretation of Bayes Theorem can be expressed as

> *For all h, $p_{t'}(h)=p_t(h|e)$, where e is the only new information available to the person between t and t'.*

Updating one's beliefs in accordance with this principle is referred to as Bayesian Conditionalisation. It is the acceptance of this principle that separates Bayesianism from other probability-based methodologies. But how can the Principle of Conditionalisation be justified?

There are two main ways in which such a justification has been attempted. One is to tackle the question directly and show that any violation of the principle amounts to behaviour that is universally acknowledged as non-rational. The second approach is to try to prove a more general result and show that Conditionalisation is entailed by that result. Let us look at these two approaches in turn, starting with the second.

The objective is to relate a person's current subjective probability function $p_t(h)$ with his future probability function $p_{t'}(h)$. Imagine that the latter could be known (or *believed* to be known) in advance. That such fore-knowledge is not uncommon, is evidenced by the pronouncements like 'I know I'll regret this, but I am going to do it anyway', 'I am sure I

will love wind-surfing', etc. Let R be a proposition expressing the person's belief, held at time t, that his subjective probability at time t' will be $p_{t'}(h)$. It has been suggested (notably, in Van Fraassen 1984) that the following condition is a requirement of rationality:

For all h, $p_{t'}(h)=p_t(h|R)$.

Van Fraassen dubbed this the *Principle of Reflection*. In stipulates that, if you know what your state of mind will be at some future point, you might as well adopt that state of mind straight away. Let us accept Reflection for argument's sake. How does it help us with Conditionalisation? Maher (1993, 121) describes the relationship along the following lines. Take any switch from $p_t(h)$ to $p_{t'}(h)$ that has been occasioned by the information acquired in the interim - to wit, by e. Since, by definition, R is a proposition describing the result of the shift from $p_t(h)$ to $p_{t'}(h)$, then $p_t(h|e)=p_t(h|R)$. The left-hand and the right-hand sides of the equation describe the causes of the shift and its result, respectively. It can be said that, given the definitions involved, the equality is analytically true.[48] Therefore, substituting this equality into the Principle of Reflection, we get $p_t(h|e)=p_{t'}(h)$, which is nothing other than the Principle of Conditionalisation. In other words, Reflection entails Conditionalisation, and accepting the former is sufficient reason for accepting the latter.

Unfortunately, for all its initial appeal, Reflection cannot be accepted as a universal principle of rationality. Maher uses the following example. Suppose that I know that, after ten drinks, I will believe myself capable of driving safely. Then, by Reflection, it would be a requirement of rationality for me to believe myself capable of driving safely after ten drinks. Since this is clearly an absurd conclusion, the example constitutes *prima facie* evidence against Reflection.

[48] The following exchange from one of the episodes of *Blackadder* may be useful. Blackadder: 'Baldrick, pass me Dr Johnson's manuscript'. Baldrick: 'You mean, pass you the thick bundle of paper that we have burned?' As defined above, 'passing the manuscript' is always 'passing the thick bundle of paper that we have burned'.

However, whereas this evidence is sufficient to apprehend the suspect, it is insufficient to convict. One can argue that the 'devil-may-care-because-I'm-drunk' example is too restrictive in scope. Scientific inferences are rarely dependent on mind-altering substances. In 'normal' contexts future belief is epistemically superior to present one because it will have come about through learning something real about the matter, whereas in Maher's example it was produced by alcohol.

In my view, this line of defence is flawed in two ways. Firstly, not all 'rational' (in some intuitive but widely accepted sense) belief shifts are occasioned by new empirical evidence. Some are caused by reflection with a small "r" - i.e., by re-thinking *old* evidence and by drawing new inferences from it.[49] Nor would it do to include 'rational reflection on the subject' in the definition of 'learning something real about the matter', for in this context there is no satisfactory way of defining 'rational reflection'. To include only deductive logic would be much too restrictive, whereas to also include probabilistic logic would beg the question, as Reflection is being defended precisely as a justification for probability theory as the logic of inter-temporal belief shifts.

Secondly, not all 'non-rational' causes result in an 'inferior' belief. The sight of the Grand Canyon, or even a relaxing drink, may well enable a person to lose the limiting preconceptions and see things in a more balanced way. The Romans did not hold that 'In vino veritas' for nothing. There are all kinds of things that, under the right circumstances, might help a person snap out of intellectual stupor. In other words, there is no well-delineated class of contexts in which Reflection becomes immune to refutation.

Moreover, there are further problems with Reflection that have nothing to do with putative counterexamples. Let us assume that at time t, my subjective probability is $p_t(h)$. Suppose that, at future times t' and t'' such that $t'<t''$, my subjective probabilities will be $p_{t'}(h)$ and $p_{t''}(h)$, such that $p_{t'}(h) \neq p_{t''}(h)$. Which of the two (or more) future subjective probabilities should we use to 'reset' the current probability function? Should

[49] Often such changes of mind result from discussing the matter with others. It is not uncommon for one juror to talk the others into changing their opinion from 'innocent' to 'guilty', or vice versa, after all the evidence had already been heard. What happens is that the evidence is presented in a different way.

it be the latest one available? But perhaps at time t' I will have access to a still more distant $p_{t'''}(h) \neq p_{t''}(h)$, such that $p_{t'}(h) = p_{t'''}(h)$, which makes $p_{t'}(h)$ superior to $p_{t''}(h)$. Should I adopt the latest probability function currently available, I would be unwittingly choosing an inferior one (unless I have perfect foresight regarding my own state of mind, which surely stretches the point too far). But unless chronological order determines epistemic seniority, there is generally no consistent way of resetting current subjective probability in accordance with Reflection.

Perhaps one would be better advised to defend Conditionalisation in its own right, rather than root it in an indefensible stronger principle. Seeing the success of the betting approach in justifying probability theory in static contexts, might it not be possible to devise a Dutch Book argument for the kinematics of partial belief? The prototypical diachronic Dutch Book argument is attributed to David Lewis. It goes as follows[50].

Let my *current* (a time t) conditional probability of h given e equal $p_t(h|e)=x$, and my current unconditional probability of e equal $p_t(e)$. Further, in the event that e occurs, let my new (at time t') *unconditional* probability of h equal $p_{t'}(h)=y$. Then it can be shown that, unless $x=y$, there is a set of bets and forfeits that leaves me with a sure loss.

Assuming that all the quotients above are subjectively fair, I should be ready to accept the following set of bargains. (i) is a conditional bet on h given e, with stake x. It pays out \$1 if both e and h occur; pays out nothing if e occurs but h does not; and is cancelled if e does not occur. (ii) is an unconditional bet on e, with stake $(x\text{-}y)p_t(e)$. It pays out $x\text{-}y$ if e occurs, nothing if it does not. Both (i) and (ii) are entered into at time t. In addition, at a later time t' (provided e has occurred), I am offered a further bargain (iii). It consists in surrendering my (now unconditional) bet on h at the price equal to my current (at time t') probability of h, that is y. Then my net gain matrix resulting from the conjunction of (i)-(iii) will look as follows:

[50] The exposition here follows Howson and Urbach (1993, p100-101).

$\neg e$	$-(x-y)p_t(e)$
e	$(x-y) - (x-y)p_t(e) - x + y = -(x-y)p_t(e)$

I lose money come what may, unless $x=y$ (in which case I break even). Therefore, goes the diachronic Dutch Book argument, Conditionalisation is the only rational rule for updating beliefs.

Unfortunately, this argument is unsound. As Maher points out,[51] losing money come what may can, under the right circumstances, be the rational thing to do. Choosing to pay insurance premiums is a case in point, as the essence of insurance is agreeing to a certain but small loss, in order to avoid an uncertain but large one.

Let us follow Maher at taking a game-theoretic look at the conditions under which sure loss is or is not a sign of irrationality. A *game* is a triple consisting of (1) a set of n players $\{p_1,...,p_n\}$; (2) the sets of strategies $\{s_1,...,s_k\}$ available to each of the n players; and (3) the set of $m=k^n$ outcomes $\{o_1,...,o_m\}$ resulting from each of the m possible combinations of the players' strategies. A 'player' is any agent capable of more than one course of action, be that a human being, social institution, computer program, animal, plant, or 'nature' in general. A 'strategy' is a sequence of 'acts', or moves (from start to finish), conducted in response to and/or in anticipation of the moves by the other players. Finally, an outcome is an n-tuple $<w_1,...,w_n>$ assigning each player i the payoff w_i at the end of the game. Naturally, w_i is determined by the strategy used by i and the strategies used by the remaining n-1 players.

For a wide variety of games, two assumptions are usually made. Firstly, it is assumed that the condition of *perfect information* holds: i.e., each player is fully aware of the possible outcomes and of the strategies open to the other players (as well as the complete history of past play). Secondly, it is assumed that not only is each player rational ("rationality"), but also that each player knows that the *others* are rational

[51] Maher 1993, 110.

("hyper-rationality"). Although a player cannot *foresee* the strategies selected by the other players, perfect information in conjunction with hyper-rationality make it possible to *calculate* them, and select his own optimum strategy accordingly.[52] Of course, the are games in which some or all of the players have more than one optimum strategy available to them ('multiple-equilibrium' games), in which case it is not possible for players to deduce exactly what the other players will do. Even then, players can calculate the full set of possible outcomes.

Such a calculation is based on each player's avoidance of *dominated strategies*. Player *i*'s strategy s_{k_i} is 'weakly dominated' if also available to him is a strategy s_{l_i} satisfying the following two conditions: (i) there is *at least one* set of possible strategies of the other players such that $w_{ki}<w_{li}$; and (ii) there is *no* set of possible strategies of the other players such that $w_{ki}>w_{li}$. In other words, a strategy is weakly dominated if there is an alternative that is better in some respects and worse in none.[53] The notion of weak dominance is extremely useful in that it gives a clear-cut and uncontroversial necessary condition for rational action, namely: 'Avoid weakly dominated strategies'.

These game-theoretical notions can shed light on the difference between the synchronic and the diachronic Dutch Book arguments. It can be demonstrated that in the latter, violating Conditionalisation need not be a weakly dominated strategy, while in the synchronic Dutch Book argument, violating probability axioms is dominated by satisfying them, as can be seen from the above examples. In a simultaneous bet on two disjoint events and against their union (with the forfeits as specified), the net gain was shown to be $r-(p+q)$. If the additivity axiom is violated, this is always less than zero for one of the parties, whereas adherence to the axiom yields a non-negative (zero) gain. Hence, in this case avoidance of sure loss is, in fact, avoidance of a dominated strategy.

---

[52] This second assumption clearly applies only to 'sentient' players, such as humans and social institutions, but arguably not to animals and definitely not to 'nature'. Nature is not 'out to get us'. Therefore, although many epistemic problems can be modelled as a game between a human player and nature, the latter's moves are not forecastable.

[53] A strategy is said to be 'strongly' dominated if there is an alternative that is better in *all* respects. However, the notion of strong dominance is less widely used in game theory since strong dominance is both less common and more difficult to prove while conferring no additional advantages in calculating players' strategies.

Let us look at the diachronic argument from this perspective. Take the example at the beginning of this section from which we started the discussion of Conditionalisation. I now believe that, should I hear the sound of thunder, I shall become 60% confident that it is currently summer. Moreover, I am prepared to stake £10, at the odds of 2/3, on such a conditional bet. If thunder does sound and it turns out to be summer, I shall end up with a £6.67 net gain. If the former occurs but the latter is not the case, I shall be £10 out of pocket. Suppose that thunder does sound within a specified period of time, and I am offered an unconditional wager *against* it being summer, with the same stake and the odds of 1/2. According to Conditionalisation, I should reject such a bet, as the odds are much lower than 3/2 (the reverse odds on the first bet). However, this time I may think to myself: "Thunder is indeed rare in winter, but it's not unheard of. On the other hand, it does feel rather chilly. I do now think summer more probable than I did before; perhaps 30%, but nowhere near the 60% I assumed before. Perhaps I was overly generous with the odds the last time around". I may then decide that, the freak stroke of thunder notwithstanding, it is probably winter after all, and that I am going to lose the £10 on my original, conditional bet.[54] On the other hand, since it will very likely turn out to be winter, I should minimise the loss and accept the new, unconditional bet, even though the odds are not as good as before. I am thinking: "The choices before me are: a 70% likely loss of £10 and a 30% chance of a £6.67 (=2/3*£10) gain, if I reject the new bet; or a 70% likely overall loss of £5 (=1/2*£10-£10) and a 30% chance of a smaller £3.33 (=2/3*£10-£10) loss, if I accept it". The second option guarantees a sure loss, but it is not dominated, under the above definition. Moreover, given the subjective probabilities assumed when throughout the example, accepting the second bet increases my expected net gain:

Option 1: $E(\text{bet}_1) = 0.7*(-10) + 0.3*6.67 = -7 +2 = -5$;

Option 2: $E(\text{bet}_1\&\text{bet}_2) = 0.7*(-5) + 0.3*3.33 = -3.5 +1 = -2.5$.

[54] That such a change of mind is not uncommon, much less irrational, was argued in relation to the Reflection principle.

Since my expected loss is halved by accepting the second bet, it is clearly the rational thing to do, even on pain of violating Conditionalisation.[55] As Maher summarises the situation,

> I suggest that the intuitive irrationality of accepting (or being willing to accept) a sure loss results from the false supposition that acceptance of a sure loss is always a dominated option, combined with the correct principle that it is irrational to accept (or be willing to accept) a dominated option (Maher 1993, 111).

We can look at the issue from a slightly different angle. After all, once e has become known for sure, $p(h|e)$ is still, formally, a conditional probability, even though it now coincides numerically with the unconditional probability $p(h)$. Conditionalisation then amount to the requirement that all conditional probabilities maintain their value through time. But there is no reason to single out conditional probabilities as immutable, in contrast to 'non-conditional' probabilities. Not only would that beg the question, but the whole distinction is nonsensical, as *any* probability can be presented as conditional on a tautology. Are we to endow with immutability only those conditional probabilities whose conditions are contingent rather than tautological? That would make no sense. Seen from this perspective, failure to conditionalise amounts to having some of our probabilities (namely, conditional ones) change over time. As Howson and Urbach suggest,

> Consider the analogy with deductive logic...my set of sentences held to be true at one and the same time is inconsistent if it contains c and $\sim c$, for some c. But of course I can consistently believe c true today and $\sim c$ true tomorrow. It is really a very obvious point (Howson and Urbach 1993, 102).

The point becomes even more obvious if we pursue the metaphor of a betting quotient as the price of a bet. Consider the price of a conventional good. I may believe a certain good worth £10 today and downgrade its subjective value to £8 tomorrow. Provided the

[55] It is easy to calculate (by substituting, for "0.7" in the equations describing Options 1 and 2, a variable and then solving for that variable) that accepting the second bet, at the odds offered, maximises expected net gain as long as the new, 'thunder-revised' probability of summer does not exceed 60%. The same result is arrived at by noticing that the original odds, based on the 60% conditional probability of summer, were subjectively fair in the sense of conferring no advantage to betting either way.

transactions are costless, I should be ready to buy it for £10 today and sell it for £8 tomorrow. Is it irrational? No, although it would be if I were willing to conduct both transactions at the same time. Who has never bought an item (a Linguaphone tape, a Hawaiian short, a set of novelty beer mats) intending to use it every day? Two days later, it is gathering dust. 'Yes, that conditional bet seemed very attractive yesterday; would you like it for half price?'

What are we left with? Clearly, Conditionalisation does not enjoy the same status as the axioms of probability. It is equally clear that there must be a wide class of cases where Conditionalisation *is* valid. What constitutes that class? The answer to that question follows from looking at what makes Conditionalisation fail where it does. As we have seen, it is the changing value of conditional probabilities. Therefore, conditionalisation holds where conditional probabilities remain invariant over time. Not only does that sound obvious, but it is also easy to prove: since $p_{t'}(e)=1$ by stipulation and $p_{t'}(h)=p_{t'}(h|e)*p_{t'}(e)$ by probability calculus, it follows that $p_{t'}(h)=p_{t'}(h|e)$. If, further, $p_t(h|e)=p_{t'}(h|e)$ (the "Invariance Condition", as it may by dubbed), we arrive at the Principle of Conditionalisation: $p_t(h|e)=p_{t'}(h)$. As Howson and Urbach put it,

> ...if, on learning e you see no reason to change your opinion that the fair betting-quotient on h were e to be true is p, say, then your only consistent value of that fair betting-quotient in the light of e's having occurred is clearly p (Howson and Urbach, 1993, 104).

We can think of our set of conditional probabilities as our "inferential structure", as they express the 'whatifs' and 'ifthens' of our mind's workings.[56] Then the domain of Conditionalisation are the situations where the inferential structure remains invariant, temporarily unaffected by the 'raw data' flowing through it. In scientific contexts, this loosely corresponds to Kuhn's periods of "normal science".

A separate restriction on the applicability of Conditionalisation arises from a formal property of Bayes Theorem, namely the fact that only positive values can change in

[56] This stipulation was also put forward by Donald Gillies as the condition of "fixity of the referential framework" (Gillies 1998, 153).

response to experience. Take the existence of God. If you absolutely rule it out to begin with (probability 0), no amount of putative miracles etc. can make your probability positive:

$$p(G|M)=p(G)*p(M|G)/p(M)$$

Well-confirmed sightings of people walking on water and feeding huge crowds with a few fish, sudden emergence of peace on earth, and other such phenomena are much likelier if God exists than if He does not, thus making the $p(M|G)/p(M)$ ratio very high. Nevertheless, that large but finite number would have to be multiplied by $p(G)=0$, thus making all the evidence epistemically impotent. That is clearly irrational; after all, a very low (but positive) probability is numerically much closer to zero than to a very high one; therefore, the first two cannot have so widely different properties as to make one of them, but not the other, immune to revision. And since in the case of zero probability such revision is impossible through conditionalisation (which transforms *prior* probabilities into *posterior* ones), it must be effected in some other way, namely buy redistributing prior probabilities in an informal fashion.[57]

EVIDENTIAL SUPPORT

Bayesianism promises a very straightforward and uniform (and moreover, quantitative) treatment of some of the key problems of confirmation theory, such as empirical support, confirmation and refutation, etc. However, it has always been criticised from a

[57] It is sometimes suggested that we should "keep an open mind" by never assigning extreme values, such as 0 and 1, to contingent propositions. Thus, even a sworn atheist should allow a positive if very low probability to the existence of God, to allow himself to be convinced by relevant experience. There are two difficulties with this proposal.

One, it is insufficiently general: sometimes we need to change our (extreme) probability assignments to analytic statements, such as mathematical theorems. If a universally accepted proof is found to be flawed, we may want to change its subjective probability from 1 to 0; the opposite is also common. If non-conditionalising change in probability is available for Fermat's Theorem, why not also for the existence of God? Thus, the problem remains, unless we erect a rigid analytic/synthetic distinction, which, post-Quine, seems too high a price to pay.

Two, the proposed solution hinges on all the possible hypotheses, including those never to be formulated, being at all times present in the domain of the probability function. Not only is it outrageous from the cognitive science point of view, but the vast majority of such hypotheses would need to have infinitely small probability (in order to fit inside the unit interval), thus again rendering them unsuited to the mathematical apparatus of conditionalisation.

variety of standpoints. Some of these criticisms appear to be misconceived; others reveal real problems. I would like to consider some of the objections and at the same time offer some of the ways in which Bayesian confirmation theory could be developed further.

Probabilification

In recent years, attempts were made to undermine the Bayesian confirmation theory, one of which was contained in an article by Peter Achinstein (1994). The theory under attack is, in a nutshell, an ambitious project of constructing confirmation theory as a probability theory (or, of explicating the notions of and relating to confirmation by means of probability functions). In his article Achinstein starts with the "probabilification thesis" (as I shall call it) that a piece of evidence strengthens a theory **iff** the theory is more probable on the assumption of that evidence (than without such an assumption):

> (1) e is evidence for h, given b, iff $p(h/e\&b) > p(h/b)$ (Achinstein 1994, 330).

Achinstein then remarks that, in a similar vein, the Bayesian proceeds to compare *different* pieces of evidence from the point of view of their relative strength:

> (3) Where e_1 and e_2 are both evidence for h, given b, e_1 is stronger evidence than is e_2 iff $p(h/e_1\&b) > p(h/e_2\&b)$ (Achinstein 1994, 331).

It is (3), which Achinstein sees as central to the Bayesian confirmation theory, that comes under his attack. He purports to show that, contrary to the Bayesian, (3) does not hold, and that greater conditional probability is neither necessary nor sufficient for greater strength of evidence, or, as he somewhat confusingly puts it[58] on the same page, that (3) provides neither a necessary nor a sufficient condition for "stronger evidence".

[58] It appears confusing because Achinstein's statement can be understood as meaning either that the second part of (3) fails to give the necessary and sufficient conditions for the first one, or that (3) taken as a whole fails to explicate the idea of stronger evidence.

Indeed, he claims that strength of evidence may decrease when conditional probability increases, and vice versa. As a consequence, the intuitive idea of strength of evidence cannot be explicated in terms of probabilities. To prove this, he argues that when the reference class that serves as evidence for a probabilistic judgement is increased, we must conclude that the evidential strength of the judgement is thereby increased also, although the resultant conditional probability may be lower than it was with a smaller reference class. I shall argue that, *pace* Achinstein, an increase in the size and/or diversity of the reference class need not (and in the general case, does not), by itself, result in strengthening the evidence.

Achinstein's first example involves drug-testing (I shall simplify it without affecting its point). Suppose a new drug was tested twice. The first study involved 1,000 people suffering from a particular illness, and resulted in 950 recoveries (evidence e_1). The second study involved also 1,000 people but only 850 recoveries (evidence e_2). What is the probability of recovery (using the new drug) in the next case? We must now adopt some inductive rule that allow us to relate the probability of a particular outcome (in our case, of recovery) in the next case to its relative frequency in the past. E.g., let us adopt the "simple inductive rule" where probability = relative frequency. Then the probability of recovery of the next patient (R) is, on evidence e_1, $p(R,e_1)=0.95$, and on the larger body of evidence $e_1\&e_2$, $p(R,e_1\&e_2)=0.90$. But although the probability of R has decreased, we feel, according to Achinstein, that $e_1\&e_2$ is stronger evidence for recovery in the next trial than e_1 alone.

The gist of this example is that as we increase the reference class, *as long as the number of favourable cases is greater than half*,[59] the results of testing become ever stronger evidence for a positive outcome of the next trial as the size of the reference class grows (I shall call this presupposition of Achinstein's "Thesis A"). However, the *probability* of a positive outcome in the next trial conditional on the previous results, may grow smaller, following the (possibly) diminishing relative frequency of positive outcomes.

[59] The stipulation that appears here in italics is not articulated by Achinstein himself, but it is necessary if his thesis is to be at all plausible.

Hence, higher conditional probability is allegedly neither necessary nor sufficient for stronger evidential support.

Despite it being neither necessary nor sufficient, Achinstein denies that increase in conditional probability is *irrelevant* for the strength of evidence. This is because *other things being equal*, including the sizes of reference classes and their representativeness (i.e., degree of variety), the degree of evidential support is indeed proportional to conditional probability (and ultimately, to the relative frequency of the positive outcome). But of course, the above condition is rarely satisfied, hence, according to Achinstein, conditional probability cannot be used to explicate the degree of evidential support, or strength of evidence.

However, this purported conclusion follows only if we accept Thesis A. Its adoption is by no means inevitable. Achinstein fails to demonstrate that the size and/or diversity of the reference class should increase the strength of evidence for a singular prediction, or, indeed, that they are at all relevant to such a prediction. (Keynes also considered such a correlation, unprompted, and rejected it.) One might undoubtedly retort that surely, size and diversity *are* relevant to something here, and I could not agree more. Just what they are relevant to, I shall suggest slightly later, in the course of assessing to what extent Achinstein's argument and his drug-testing example justify his conclusion. For the moment let us merely note that if there are ways, other than Thesis A, to accommodate the importance of size and/or diversity considerations with respect to evidence, Achinstein's main argument collapses.

First I shall consider the extra twist in Achinstein's argument which hinges on his Michael Jordan example. In order to further divorce strength of evidence from conditional probability, Achinstein introduces another (besides the reference class size and variety) consideration that affects the former: *explanatory connection*. Says Achinstein:

> An explanatory connection between evidence and hypothesis is another factor that can strengthen evidence without increasing probability (Achinstein 1994, 337).

The example that Achinstein gives involves the following three statements:

b: No man has ever become pregnant.

e: Michael Jordan eats the breakfast cereal Wheaties.

h: Michael Jordan will not become pregnant.

Achinstein rightly notes that *e* is not evidence for *h*.[60] His explanation of this fact is that there is no explanatory connection between them, which is fair enough. However, the conclusion he draws is puzzling. It is, once again, that conditional probability does not explicate the strength of evidence, for $p(h, b\&e)=1$ ("approximately", says Achinstein, as if in doubt), despite the fact that *e*, by itself, is not evidence for *h*. What immediately strikes one is that whereas the conditional probability he considers is conditional on *b&e,* the piece of evidence whose strength he (rightly) denies is *e* alone. This shows the limits beyond which one cannot consistently employ ill-defined notions like "strength of evidence". Probabilities are used to explicate a number of different concepts having to do with confirmation. There are the well-established notions of *degree of confirmation* defined, in the Bayesian tradition (for example, by Carnap), as

$$c(h, e\&k)=p(h, e\&k)^{61},$$

and that of *weight of evidence* defined as

$$w(e, h, k)=p(h, e\&k)-p(h, k),$$

where *k* is background knowledge. Simple inspection shows that the two notions are substantively different. Whereas degree of confirmation is a 2-place function, weight of evidence (also referred to as 'support') is a 3-place function.[62] This is not surprising,

[60] Even if *e* included a reference to Michael Jordan's gender, it would not be evidence for *h* by itself, i.e., without a background assumption that men don't (or can't) become pregnant.

[61] Of course, one may well protest, following Popper, that confirmation is *not* a probability function; but Achinstein's argument consists exactly in taking certain probabilistic ideas for granted and trying to show that they lead to counter-intuitive results. Rejecting '$c(h, e\&k)=p(h, e\&k)$' on other grounds would amount to skipping the argument altogether and proceeding directly to the conclusion.

[62] The notion of weight of evidence, $w(h, e, k)=p(h, e\&k)-p(h, k)$, was first introduced by Alan Turing and I. J. Good in the context of their code-breaking work during WWII. It is discussed at length by

since the former measures the overall empirical standing of a hypothesis, while the latter singles out the contribution of a specific piece of evidence.

Now it is true that the degree of confirmation conditional on *b&e* (together with the tacit assumption that Michael Jordan is male) equals 1; but by the same token, one cannot deny that *b&e* (plus the tacit assumption) is strong evidence for *h*. It is just that all the strength of this evidence is concentrated in the first premise, *b* (plus the tacit assumption), whereas *e* does not add anything. This is nicely matched by the (probabilistic) fact that

$$w(e, h, k)=p(h, e\&k)-p(h, k)=1-1=0,$$

where *b* is a part of *k*. (Alternatively, one can just substitute *b* for *k* throughout this last formula for the case where relevant background knowledge, besides *b* and the tacit assumption, is null.) To sum up, if the Jordan example is supposed to turn on the notion of degree of confirmation, then it "works" only because Achinstein illegitimately oscillates between *e&b* and *e*; but if Achinstein means "weight of evidence", then the example does not work at all and, on the contrary, confirms the "probabilistic" intuitions while disconfirming his claim that explanatory connection has nothing to do with probability.

The unfortunate Jordan example aside, one cannot help agreeing that sample size and/or variety *are*, somehow, relevant in evaluating empirical evidence. How exactly can these factors be accommodated probabilistically? One of the ways to do this would be to construct a confirmation function that always increases and decreases together with our intuitive feeling of the strength of evidence (the latter being based, among other things, on sample size and variety). A number of further examples devised be Achinstein show that various attempts to do so (particularly, in Carnap 1950) have failed; a similar claim (albeit not substantiated in the article) is made with respect to a number of more recent proposals in the literature. However, the claim that all known proposals of devising an

---

Donald Gillies in his (1998), where the author also develops his own theory of confirmation, which is neither Bayesian nor orthodox Popperian.

"all-embracing" confirmation function whose behaviour would closely follow that of the intuitive "strength of evidence" feeling have failed, prompts at least two responses.

The first one is that past failures do not preclude a success in the future; in fact, human ingenuity consists, to a large extent, in the ability to turn failure into success. Why could it not be so with devising "all-embracing" confirmation functions? Among recent proposals how to do so is Donald Gillies' "testing measure of confirmation".[63] His theory is based on two principles. The first one is Popper's "principle of severe testing" which in Gillies' paraphrase states that

> ...the more severe the tests which a hypothesis h has passed, the greater is the confirmation of h (Gillies 1998, 157).

The second principle is that of "explanatory surplus":

> The principle denies that if e follows logically from h, this automatically means that e supports h. Not all the facts which follow from a given hypothesis support that hypothesis, so the principle claims, but only a subset of these deducible facts – a subset which constitutes an explanatory surplus (Gillies 1998, 160).

Gillies employs Popper's measure of the severity of a test, $Q(h,e_i,b)=p(e_i \mid h\&b)-p(e_i \mid b)$, where $1 \leq i \leq n$. This measure is calculated for each i such that e_i is part of h's "explanatory surplus". The positive and negative values are then added up separately, resulting in Q^+ and Q^-, respectively. The ordered pair $<Q^+,Q^->$ expresses the confirmation of h by $e_1,\ldots, e_n$.

Gillies' degree of confirmation is clearly not a probability, by virtue of being an ordered pair, and is therefore not suitable for Bayesian confirmation theory. However, it is a plausible explication of the intuitive idea of the strength of evidence, and as such represents a counter-example to Achinstein's contention that an "all-embracing" confirmation function cannot, in principle, explicate the pre-theoretic idea of the

[63] While Donald Gillies modestly denies his confirmation function being 'all-embracing', and indeed doubts that any one confirmation function will fit every possible application, his proposal is certainly general enough to illustrate my point.

strength of evidence. If a neo-Popperian function of this kind is possible, why not a neo-Bayesian one?

Even more importantly, in my view, "all-embracing" confirmation functions need not perfectly reflect the behaviour of our gut feelings about evidence. Demanding otherwise, as Achinstein does, is analogous to dismissing the scientific notion of temperature on the grounds that the temperature function does not behave exactly as our feeling of warmth and coldness; indeed, a metal surface (cold or hot) produces a stronger corresponding sensation in the skin than a wooden surface, even though the temperature of the latter may be more extreme. So, the second possible objection to Achinstein is that he does not appreciate why and how we go beyond (and contradict) intuition in the first place.

But my ultimate sympathy is with the "two-tier probability" approach that Achinstein considers (and rejects) in the last section of his article. According to this approach, the strength of evidence for an *outcome* is an increasing function of relative frequencies (and possibly other factors such as Carnap's "relative widths" which are available, of course, only in formalised languages) and equals the probability of that outcome conditional on the arguments of such a function. Next, the probability (degree of confirmation) of a first-order *probabilistic hypothesis* is a function of other factors, such as the degree of variety of the reference class, its size, etc. (cf. Carnap 1950). The difference between the first-order degree of confirmation of a singular prediction and the second-order degree of confirmation of the first-order hypothesis is, essentially, the difference between the *strength* of the evidence for a prediction and its *reliability*. We thereby supplant the "naive", pre-analytical feeling of strength of evidence with two "scientific" notions: of strength, and of reliability. Analogously, the pre-scientific feeling of warmth/coldness of various objects was supplanted (in scientific discourse) by the notions of *temperature* on the one hand, and of *heat capacity* on the other; not surprisingly, neither of those two notions behaves as the pre-scientific feeling that they were brought in to explicate.

This approach also allows one to solve the so-called "paradox of ideal evidence" introduced by Popper to undermine epistemic interpretations of probability (i.e., those

treating probability as a characteristic of uncertain knowledge rather than as an objective magnitude). The paradox goes as follows. Suppose we have a coin which *looks* symmetrical. On most epistemic interpretations, we will conclude that the probability of "heads" in a certain toss equals 0.5, since there are no grounds to conclude that "heads" is more probable than "tails", or vice versa. Suppose further that we toss the coin 1,000 (or as many as we like) times, and the relative frequency of "heads" approximately equals that of "tails" and 0.5. The paradox, according to Popper, consists in the fact the 1,000 tosses constitute overwhelming ("ideal") evidence in support of the conjecture that p("heads")=0.5 and hence have dramatically altered the epistemic situation, while according to the position he criticises, these tosses must be irrelevant, for p("heads") has not changed. This paradox can now be dismissed on the grounds of the inference

'p("heads") has not changed'
$\therefore$ 'The tosses are irrelevant'

being invalid. It is invalid because it rests on a false tacit premise - namely, that the probability of a certain outcome of the next toss is *all* that there is under consideration. However, apart from the hypothesis relating to the next toss (characterised by the value of p("heads") that we venture), we are faced with another one relating to the degree to which the toss is *typical* of the long-term behaviour of the coin. It is the latter hypothesis whose assessment involves the size and variety of the reference class (in this case, the 1,000-strong class of "trial" tosses).

This deconstruction of Popper's argument need not be viewed as fully rejecting his own theory of probability. The first tier in the two-tier approach may well be interpreted, following Popper, objectively, e.g., as the propensity p to produce the outcome R (in Achinstein's drug-testing example, recovery) which we assign to a particular experimental set-up S (in this example, an organism infected in a particular way and treated with a particular drug). This propensity characterises, possibly indirectly, the strength of evidence for R in any particular case generated by S. Note that p does not depend on the number of instantiations of S or on their diversity; what does, arguably, depend on them, is the degree of confidence that the value of p we have chosen is the right one. This second-order degree of confidence is, essentially, Keynes's "weight of

evidence"[64] (with the addition of considerations of diversity). It is at this level that the size of the reference class matters, not at the level of choosing the value of p.

What has Achinstein got against this approach? He presents two objections that are difficult to reconcile:

(1) in the case of second-order probabilities, different aspects of the evidence (e.g., its variety and sample size) may still "pull us in different directions", whereas any attempt to find a "vector sum" might be counter-intuitive (in the sense that the "vector sum" increases when the feeling of strength of evidence decreases, and vice versa); and

(2) nobody seriously pursues the two-tier approach anyway.

As regards (2), in my view it would not be a serious objection even if it were true, and it is not, either.[65] And as far as (1) is concerned, I reiterate the counter-arguments outlined at the end of section **3.**

To sum up, the intuitive idea of strength of evidence should be analysed into the separate notions (1) of the degree of confirmation conditional on evidence (possibly defined by means of objective probabilities, such as relative frequencies or propensities), and (2) of the degree of reliability of that evidence (this latter also being a conditional probability function).[66] This separation (instead of the previous unity of the corresponding gut feeling), as well as the (putative) fact that at the level of reliability assessment, the probability function will sometimes behave counter-intuitively, merely show that confirmation theory is not a simple extension of common sense; but they

[64] Not to be confused with the Turing-Good concept of the same name. By 'weight of evidence', Keynes meant the sheer quantity of evidence available, and therefore the reliability of inferences based on that evidence.

[65] The whole of Bayesian statistics revolves around calculating, given the evidence, the probability of statistical hypotheses which themselves, by definition, include a statement of probability. The body of literature devoted to this problem is very considerable indeed.

[66] Keynes himself did not believe weight of evidence to be a probability function. I think it is an open question, but my feeling is that it can be construed as one.

cannot, *pace* Achinstein, serve as arguments against explicating the feeling of "strength of evidence" in terms of conditional probabilities.

Synthetic universal statements

"Zero probability"?

In the previous section we were concerned with singular predictions. But the most interesting kind of inductive inference is universal generalisation. Whether or not the Bayesian confirmation theory is able to handle those is the touchstone of this theory. It is well known that, due to the form of Bayes's theorem[67], a statement can receive positive confirmation only if its prior probability is positive. The opponents of Bayesianism (notably, the Popperians) insist that the latter fails, if not for any other reasons, because non-tautologous, or synthetic, universal statements (and hence all universal scientific hypotheses) have zero probability (if the domain is infinite, which will be assumed from now on). Popper himself used this claim to argue that no confirmation theory (which assigns some synthetic universal statements positive degrees of confirmation) could be constructed probabilistically, as an inductive logic. There are *a priori* reasons to doubt this claim; for instance, why should all scientific theories be no more probable than logical contradictions?

Popper's claim, made in Appendix *vii of his *Logik der Forschung* (1934, expanded English edition 1959), is that non-zero probability ascriptions to synthetic universal statements (interpreted on an infinite domain) must be inconsistent. This is discussed and rejected in several articles by Colin Howson.[68] Howson shows that Popper's argumentation is based on a number of logical errors. Popper's putative proof falls into three separate arguments which Howson calls (a) the argument from independence; (b) the argument against Jeffreys; and (c) the 'dimensional' argument'.

---

[67] As explained above, Bayes Theorem relates the posterior probability of a hypothesis (that which is conditional on certain evidence plus background knowledge) to its prior probability (which is conditional on background knowledge alone), the prior probability of the evidence itself, and its probability zconditional on the hypothesis (the likelihood):

$p(h, e\&b)=p(h, b)p(e, h\&b)/p(e, b)$.

[68] Criticism of Popper's argumentation can be found in Colin Howson's (1973) and (1987).

(a) Popper's 'argument from independence' goes as follows. Take a universe with a single predicate F and n individuals c_1, ..., c_n. Then there will be j statements of the form u_{ij}=Fc_j or $\neg Fc_j$, where $1 \leq j \leq n$ and $1 \leq i \leq 2^n$. Each 'possible world' (not Popper's term) within that universe can be described as $\mathbf{u}_i$, an n-long conjunction of u_{ij}. Clearly, there will be 2^n of these. Assuming that the u_{ij} are independent (each one having probability 1/2), all the $\mathbf{u}_i$ are equiprobable. Once the assumption of a uniform probability distribution over the $\mathbf{u}_i$ is granted, Laplace's 'classical' measure gives each one probability 2^{-n}. Since the universal statement $\forall xFx$ is equivalent to one of the $\mathbf{u}_i$, namely that one which does not contain any "$\neg$"s, its probability is also 2^{-n}, which tends to 0 as n tends to infinity.

Of course, this argument lives and dies with the assumption of independence.[69] Popper gives no real justification except to say that we have no reason to assume otherwise:

> ...the only reasonable assumption seems to be that ... we must consider [the elements u_{ij}] as mutually independent of one another ... every other assumption would amount to postulating *ad hoc* a kind of after-effect (Popper 1959i, 367).

To this Howson replies, quite naturally, that postulating a *lack* of every possible after-effect is just as *ad hoc*, it amounts to making just as substantive a claim about the universe and does not have, to use Popper's expression, 'the character of a tautology, valid in every possible universe' (which, according to Popper, is the *sine qua non* of a proper theory of logical probability):

> In other words to adopt the assumption of Popper's arguments is to commence with an extreme bias against the possibility that generalisations of any type will hold. [...This] is both arbitrary and, given the evidence which strongly suggests that the universe is a highly structured place, perverse (Howson 1987, 213).

A further problem with Popper's assumption of a uniform prior distribution is that no probability distribution in more than one parameter is genuinely uniform:

[69] Or, alternatively speaking, the assumption of a uniform probability distribution over the $\mathbf{u}_i$.

> In the first place, no uniformly distributed measure over an infinite or semi-infinite interval of Euclidean *n*-dimensional space can be normalised, so there can be no uniform probability distribution over the unrestricted parameter space of a theory with ***n*** free parameters. Secondly, even if the parameter values are restricted to a bounded region, any nonlinear reparametrisation will destroy the uniformity of the measure, which is the main reason why historically uniform prior measures ceased to be regarded as an integral part of Bayesian methodology (Howson 1987, 212).[70]

Since Popper's argument cannot even be stated consistently, *a fortiori* it is invalid.

(b) Popper's argument 'against Jeffreys' involves the following result derived in Jeffreys' *Theory of Probability* (1939). Let a be a theory entailing denumerably many instantiations $b_1,\ldots,b_n, \ldots$, and let $b^n=b_1\wedge\ldots\wedge b_n$. Then either $p(a)=0$, or $p(b_n \mid b^{n-1})\rightarrow 1$.[71] Popper argues that the second alternative is inconsistent, therefore it must be the case that $p(a)=0$. Whence the inconsistency?

Take any law a. For each n, we can construct a law a_n stipulating that for the first n-1, the property exemplified by a holds, but for n and all subsequent instantiations it does not.[72] Let us assume, says Popper, that Jeffreys' convergence holds, that is

$$p(b_n|b^{n-1})\rightarrow 1.$$

For some n, $p(b_n \mid b^{n-1})$ must be close to 1. But a_n is also instantiated by b^{n-1}, not just a. Hence, $p(\neg b_n \mid b^{n-1})$ must also be close to 1, which violates the probability calculus.

However, Popper's *reductio ad absurdum* crucially depends on the assumption that 'there will always be predicates A and B which both apply to all things so far observed,

---

[70] Howson is referring to the various 'geometric paradoxes' that plagued Keynes' Principle of Indifference.

[71] The second alternative is, of course, nothing other than a 'limit' reformulation of Laplace's Rule of Succession.

[72] Howson labels such laws 'Goodmanesque', referring to Nelson Goodman's famous examples of 'bent' predicates such as 'grue' (green up to the *n*th trial and blue thereafter) and 'bleen' (vice versa).

but lead to incompatible probabilistic predictions with respect to the next thing' (Popper 1957, 371). Otherwise, the argument collapses.

Howson, in his (1973), shows that such predicates exist, far from 'always', only under a very special condition. Take a series of 'bent' laws a_k where the 'break point' occurs at the kth instantiation (the original law a can be viewed as the limiting case, a_0). For each k, define the instances of a_k thus:

$c_{ik}=b_i$ for $0\leq i< k$;

$c_{ik}=\neg b_i$ for $k\leq i$.

Then Jeffreys' convergence can therefore be re-stated thus:

As $j\rightarrow\infty$, $p(c_{ik} \mid c^{(j-1)k})\rightarrow 1$.

Now let us consider $p(c_{ik} \mid c^{(j-1)k})$ as a function of k:

$f_j(k)=_{def} p(c_{ik} \mid c^{(j-1)k})$.

Howson demonstrates that Jeffreys' convergence leads to a contradiction only on the condition that $f_j(k)$ converges uniformly to 1 over the set of $k\geq 0$.[73] If this condition does not hold, Popper's *reductio* falls through. Therefore, all that follows from Popper's argument is that non-zero probability of universal hypotheses requires non-uniform convergence, not that non-zero probability is inconsistent.

(c) Popper's 'dimensional' argument has to do with the ordering of general theories by their simplicity. It consists of two parts: a lemma purporting to prove that more complex theories are *a priori* more probable, and a simple inference from that. Assume that the lemma is true; then for each theory of complexity n there is a more probable one of complexity $n+1$, etc. If one of the theories in the sequence has positive probability, then the countable sum of their probabilities equals infinity, which is absurd.

[73] This condition favours those 'bent laws' whose bend is further down the sequence and which are,

Of course, none of this follows, since Popper's lemma is fallacious. It hinges on the classical idea of 'counting' the possibilities favourable to a theory. Take a theory T^d of dimension d; it will be represented by a subset of the Euclidean parameter space of d dimensions R^d. Let a 'cube' of side ε represent a region within R^d that is consistent with the theory, that is, a possibility favourable to it. Next, says Popper, look at a theory T^k of dimension d ($k<d$) which is obtained from T^d by dropping d-k adjustable parameters. Topologically, T^k is a projection of T^d into the space R^k. Since R^k is 'smaller' than R^d, goes the argument, every one of its regions will contain fewer 'cubes' of side ε than the corresponding region of R^d, and therefore fewer possibilities favourable to T^k than the corresponding region of R^d contains possibilities favourable to T^d. Hence, concludes Popper, $p(T^k)< p(T^d)$.

There are several ways to discredit this line of reasoning; I propose the following, which to me appears the most obvious. What is the dimension of Popper's 'cube' of side ε? It is d in R^d but k in R^k. Every cube ε^k in R^k corresponds to a continuum of cubes ε^d in R^d. Therefore, each possibility favourable to T^k subsumes a continuum of possibilities favourable to T^d. Of course, this does not imply that T^k is infinitely more probable than T^d, since it also has an additional continuous infinity of *unfavourable* possibilities to deal with. In general, there is no way of telling whether theory T wins or loses when some adjustable parameters are dropped, since the metric may be different in the two spaces. However, to illustrate Popper's fallacy, picture a 3-dimensional space and its 2-dimensional projection. Suppose that each of the 3 adjustable parameters of theory T^3 ranges between 0 and 4, and that T^3 is true only if each parameter does not exceed 3. Then the overall domain of T^3 has volume 4x4x4=64, while the region favourable to T^3 has volume 3x3x3=27. This gives T^3 prior probability $27/64\approx0.42$. Meanwhile, T^3's projection into the 2-dimensional parameter space enjoys the 'favourable' volume of 3x3=9 out of the overall volume 4x4=16, giving it prior probability $9/16\approx0.56$. Dropping a parameter may, *pace* Popper, increase prior probability.

Thus, none of Popper's arguments achieves its stated purpose of proving the zero-probability thesis. That Popper is misguided in his attempts, is illustrated by the fact

---

therefore, more similar to the corresponding 'straight law'.

that, whereas Carnap's *c*-functions are, indeed, such that all synthetic universal statements receive zero probability, there are other systems, e.g. Hintikka's, where this is not the case. But even if Popper's arguments were sound, they would only apply to the so-called logical probability function which, when all the adjustable parameters have been fixed, is uniquely determined for any pair of arguments (and I have cited one of the reasons why logical probability is a fiction). In subjective interpretations of the probability function, the assumption of uniqueness does not hold and zero-probability prior ascriptions turn out to be just as arbitrary as any others.

This does not mean that the subjectivist approach does not have any difficulties in this regard. Some claim that the subjectivist also has to assign zero probability to all synthetic universal statements, although for a different reason.[74] According to the orthodox subjectivist interpretation, synthetic universal statements cannot have positive probability because they cannot be consistently bet on at positive odds and thus do not warrant a positive degree of belief: while the bettor-against can win (if he can produce a refuting counter-instance), he cannot lose (since no number of confirming instances will be sufficient to prove a genuinely universal synthetic statement). This, supposedly, violates the requirement of strong coherence according to which to be rational, a set of betting quotients must rule out the situations where the bettor-on wins in none of the possible cases and loses in at least one of them.

According to Gillies, this alleged flaw does not refute Bayesianism completely but confines it to those contexts where "all the hypotheses involved are singular statements rather than universal scientific laws" (Gillies 1998, 155). For instance, "it seems to me that there is a good case for saying that Bayesianism is applicable to legal reasoning" (Gillies 1998, 156). This to me looks suspiciously like damning with faint praise, and I shall try to show that restricting subjective Bayesianism to singular statements is quite unnecessary. But first, a technical remark.

[74] The following argument dates back to the early 1970s and is stated in, e.g., L. J. Cohen's *The Implications of Induction* (1970) and M. B. Hesse's *The Structure of Scientific Inference* (1974). More recently, the argument has been restated by Gillies in his (1998).

Even if the "no-win" argument were as compelling as it seems, synthetic universal statements need not have *zero* probability; the probability function may just not be defined for them (in fact, this seems to be the most plausible reading, shared e.g. by Ramsey). Rather than insisting on a zero betting quotient, the rational punter would decline the bet altogether. This is scant consolation to the Bayesian, since if synthetic universal statements had either zero probability or *no* probability, then the notion of empirical support for such statements would be unintelligible, after all. However, there are a number of ways in which this difficulty can be overcome.

Howson (1987) and Howson and Urbach (1993) tackle the problem by denying that degrees of subjective probability are identical with (rather than measured, quite fallibly, by) betting quotients. A subject's personal probability assigned to a hypothesis *h* is the quotient he would find fair *if* there were a way of eventually determining *h*'s truth value. When eliciting personal probabilities, Howson and Urbach in effect ask the subject to assess a counter-factual situation and hence - to exercise his imagination, to a degree higher than that which is normally involved in forming a non-hypothetical judgement. Unlike the more traditional betting approach, this precludes the possibility that for some meaningful statements, personal probability is not defined.

In my view, this approach, though theoretically absolutely correct, has a pragmatic drawback in that in often does not yield a unique numerical value. For instance, if I were asked what, in my opinion, constituted the fair betting quotient on the discovery of a new planet in the Solar System within the next year, I might say "1,000/1" today and "500/1" tomorrow, with no change in my background knowledge. On the other hand, if I were required to back up my judgement with hard cash, I would likely be less liberal in my judgements. A suitable incentive really concentrates the mind.

In general, there is a wide range of circumstances in which the subject will produce a range of values that he deems to be fair quotients with which he would be able to bet on or against the hypothesis in question. Or, to put it differently, the lowest quotient with which he would be willing to bet on the hypothesis may not coincide with the highest quotient with which he would be willing to bet against it, the interval between them representing the subject's degree of belief in *h*.

It is well known that interval-value interpretations of probability do not enjoy the strength of point-value interpretations (since most of the equalities in the calculus of probabilities become inequalities). Hence, it is desirable to keep point values. The Howson-Urbach approach does not, in general, allow that, for it provides no mechanism whereby the subject would be compelled to give a number rather than an interval.[75] So, if we abandon actual bets, in at least some cases we are stuck with weaker calculus; but if we keep them, supposedly we cannot handle synthetic universal statements.[76]

I would like to argue, however, that synthetic universal statements *can* be consistently bet on - not because synthetic universal statements *can* be *proven* by a finite number of supporting instances (they cannot), but because they *cannot* be *proven false* by a finite number of counter-instances, either. The rationale for this latter belief is *epistemological holism* which is fairly uncontroversial (unlike *semantic* holism) in modern analytic philosophy. More specifically, my claim has to do with the fine mechanics of betting and settling bets.

A betting interpretation for universals

In all betting-based approaches in probability theory, there must be a mechanism or procedure for deciding the outcome of a bet. Let us ask ourselves how the outcome could be ascertained. It is easy enough when the event on which the bet was made is describable by an observation sentence (provided both parties belong to the same linguistic community or if not, the sentences that describe the outcome are observational

[75] Howson takes issue with this claim. In private correspondence he suggests that the stronger calculus can still be maintained on pragmatic grounds. After all, goes his argument, "very few quantities in science are truly continuous, but that doesn't stop people using continuous mathematics". I find the analogy a valid one; however, for it to work, the intervals must be small. Then we could, e.g., take the middle point and call it the individual's point-value probability. But this simply shifts the original problem: in cases where the original interval-valued probability estimation is too wide, how do we compel the individual to narrow it down to the extent that it can be safely represented by a point?

[76] There is a further objection to the latter option, namely that the betting situation itself may influence the judgement of probability, leading to inaccurate measurements of the degree of belief. However, it is not at all clear why this should generally be the case. If the stake is kept sufficiently small compared with the subject's total wealth, and any extraneous considerations are excluded (such as Maher's bus fare scenario described above), then the betting arrangement is non-distortive. Even the subject's attitude to risk does not materially affect his probability judgement. Suppose one is risk-loving and is therefore willing to accept shorter odds than what would be strictly reasonable. But by stipulation, this would amount to being over-cautious on the associated opposite bet. The reverse applies to risk-averse individuals. As a result, irrespective of the attitude to risk, all individuals will tend to offer 'risk-neutral' fair odds.

in both languages). By Quine's (mark 1 - *Word and Object*) definition of an observation sentence, which I accept, the parties are bound to agree about the outcome. An example of that is the outcome of a race when the finish line is in plain view of both parties and the contestants finish with a gap of, say, 3 feet. If either of these conditions is not satisfied, the parties defer to the announcer who delivers the official verdict. Even when the parties to the bet could see for themselves which horse came first, their observation may be overruled by the judges if they find irregularities in the way the race was conducted, etc. In a different way, the parties defer in their determination of the outcome to technology and organisation when they are watching the race on TV. Except for the simplest cases, the parties to a bet will agree about its outcome only if there is a social mechanism whereby a determination is imposed. Putnam, in "The Meaning of 'Meaning'", argued that community-wide standards of meaning hinge on the division of linguistic labour and on the speakers' deferring in their use of certain words (such as natural kind words) to the relevant community of specialists. I believe that, analogously, the parties' ability to settle a bet hinges on the division of authority and their deferring to the relevant institute of authority (be that state, science, religion, sports federation, or mass media).

The point I am trying to make is that even in straightforward cases like those that the frequency interpretation considers its own, the possibility of consistent and "fair" (in however weak a sense) betting depends on the availability of a procedure whereby all the parties defer in their determination to a third authority and whereby any possible disagreement may be effectively settled. Why then could not a fairly natural procedure be devised whereby a positive outcome of a bet on a synthetic universal statement could be ascertained?

Falsifiability

Now the onus is on me to show that such a procedure is possible. What is it to say that a certain general hypothesis is true? It implies that there will be no counter-instances. How long do we have to wait before we despair of finding a counter-instance? An analogy with other situations where claims of similar strength are considered "virtually proven" (despite similar lack of conclusive evidence) might be helpful. In US law, for instance, a person is considered legally dead if, despite a thorough search, there is no

evidence of his/her existence after seven years. Analogously, the parties to a bet on a synthetic universal statement could agree ahead of time what amount of experimental testing they deem sufficient for its positive confirmation. (Incidentally, the parties also need to agree what amount of *negative* evidence they deem sufficient for its *refutation*; clearly we don't want to abandon a promising theory after one report of alleged refutation.) In fact, the parties conclude a *deferred* bet, with the understanding that the payoff will be postponed until after the evidence is in which, to their mutual satisfaction, determines the truth value of *h*.

It might be objected that no such procedure exists for general theories of e.g. physics. I would counter that, firstly, "betting on theories" need not be explicit, and secondly, although there is no universally agreed procedure identical for all scientific contexts, the opposite parties can usually agree on what new evidence would settle their differences in a specific quandary. Taking a cue from Newton, such evidence is often referred to as the "crucial experiment". I do not intend to go into any problems that may exist with this notion, except for stating that, rightly or wrongly, it does have currency in scientific discourse.

A deferred bet is *not* a conditional bet, since it would not be called off no matter what evidence comes in. Granted, any procedure devised for settling a deferred bet involves a certain amount of arbitrariness; but not more than the amount involved in other methodological decisions, such as determining the boundaries of an outcome distribution beyond which a statistical hypothesis is considered falsified.

The somewhat conventional character of the stipulations I mentioned should not be misrepresented as external to probability theory. On the contrary, the rigorousness of such stipulations can be used to measure the *degree of reliability* of our probabilistic judgements.[77] (I believe that in general, second-order probabilities measure nothing other than the degree of reliability of our first-order probabilistic judgements, as I urge in Chapter I.) For instance, the betting situation wherein the parties agree that a

[77] Of course, whether such numerical second-order measure functions are possible and if so, how they could be constructed, is a separate question.

hypothesis is refuted after one counter-instance and vindicated after ten confirming instances, would not yield a very reliable prior assignment.

I would like to emphasise that I am not proposing *conditionalisation* in the guise of a priori assignment. The evidence needed to settle such a bet is *not* used to alter a probability. In fact, nothing would change if neither positive nor negative evidence ever came in at all: the bet would have to be suspended or called off (if the appropriate provisions were in place), but the prior probability would have already been elicited.

"Zero probability" revisited

I do not wish to argue, however, that *all* synthetic universal statements can have a positive probability. The Bayesian will agree that some classes of universal scientific statements do have zero probability, namely, deliberate *idealisations*. As an example of such an *a priori* false generalisation, one can take the localisation of finite mass in a mathematical point in Newton's mechanics, or constant prices in some economic models; etc. Such universals are *born* with zero probability.

In general, zero prior ascriptions are justified only to those synthetic universal statements that are extremely improbable in the light of background knowledge. Normally, a zero-probability ascription is followed by a very low ranking of the statement within scientific discourse. The major exception is the class of those statements which are admitted as deliberate idealisations - false (by default) but convenient. In their case, the only justified prior (and posterior) probability is indeed zero, but they are not admitted as potential truths, hence empirical confirmation procedures are not applicable to them in the first place. Thus, zero probability of universals results, where it does, from the peculiar function of deliberate idealisations within scientific discourse and not, *pace* Popper, from any general properties of universal statements.

The class of synthetic universal statements to which Popper's zero-probability claim is applicable is, not accidentally, the class of statements on which consistent bets cannot be made (even with the amendments I propose above), for their truth value is known ahead of time. In general, however, the subjectivist version of Bayesianism indicates the

ways in which non-zero probability can be consistently ascribed to synthetic universal statements - either the hypothetical Howson/Urbach approach, or, where it fails deliver a point value, the "acceptability-based" approach outlined above.

Conclusion

The two sections of this chapter dealt, respectively, with the epistemological applications of the Bayesian confirmation theory and with its technical underpinnings My objective was to show that this theory does apply across the board, *pace* Carnap and Popper who contended that it could not be extended to universal generalisations which are, as it is now universally recognised, the central part of science, and to combat the claims that this theory leads to epistemologically absurd consequences.

CHAPTER 3: SIMPLICITY

The criticism levelled against the Bayesian theory of confirmation and inference by Popper and his followers (such as David Miller) was, to put it bluntly, that such a theory was conceptually *impossible*. However, nowadays few go that far; the critics allege merely that Bayesianism is *unworkable*, for one reason or another. Achinstein's attempt to show that Bayesian procedures do not do justice to the scientific notion of confirmation belongs to this category. Following a similar strategy, in a number of recent articles[78] Malcolm Forster and Elliott Sober make a case against the Bayesian understanding of the role of simplicity in scientific inference:

> The central problem with Bayesian philosophy of science is that it cannot take account of the relevance of simplicity and unification to confirmation, induction, and scientific inference (Forster 1995, 399).

The authors dismiss as a myth, among other things, "the standard Bayesian folklore about factoring simplicity into the priors"[79]. Their case is based on his analysis of the procedure of presenting one magnitude as a function of another (in the statisticians' jargon, "curve-fitting"). In the above mentioned article, Forster argues that the notion of simplicity that emerges in the course of examining this procedure speaks against the Bayesian understanding of scientific rationality. I critically examine his and Sober's argument and show that at most, it undermines an unnecessarily restricted version of Bayesianism but not its core principles.

In Section 1, I state Forster and Sober's claims based on their assessment of certain developments in mathematical statistics, and try to make explicit the background assumptions, some of which are never stated by the authors in question, without which their claims cannot possibly make any sense. In Section 2, I critically assess the purported significance of their methodological claims for Bayesian epistemology and philosophy of science, and argue that even if the authors' tacit and overt assumption had been warranted, their conclusions would still not have validly followed from them. Those assumptions are subjected to scrutiny both during my exposition in Section 1 and in the course of offering, in Section 3, a positive alternative to Forster and Sober's theses.

---

[78] Forster and Sober (1994), Forster (1995) and (1995i), Forster and Sober (1999).

[79] Forster (1995), 399

THE EMPIRICAL ADVANTAGES OF A SIMPLICITY CRITERION

In their 1994 article, Forster and Sober present the results of the Japanese statistician H. Akaike[80] who, they claim, showed how simplicity could be explicated in relation to verisimilitude, or closeness to the truth. From that exposition they subsequently draw a number of methodological conclusions.

The background to Akaike's results is as follows. Suppose we want to present an observable magnitude (the "endogenous variable" in economics, or the "dependent variable" in statistics) as a function of another (possibly observable) magnitude (the "exogenous variable", or the "independent variable", respectively). To that end, statisticians take as their data a collection of observed points (pairs of particular values of the two variables) and try to find the curve that expresses the true relationship between the two variables. In doing so, the standard assumption is that, subject to certain conditions laid down below, the curve "closest" to the "true" one is that which "fits" the observed data the "best."[81] Akaike certainly accepts this assumption, as will be seen from his definition of distance from the truth (the latter is defined by him in terms of goodness of fit). But innocuous as it may seem, this assumption involves a number of notions that, to be acceptable, require very careful explication. To that end, in my view one needs to answer the following three questions:

(1) What is the appropriate measure of degree of fit between a curve and the data?

(2) What is the "true" curve?

(3) What is the distance from an arbitrary curve to the true one and how can it be measured?

Goodness of fit

The first question is also the one that is given the clearest answer. Traditionally, statisticians measure degree of fit by the "(corrected) sum of squares" (SOS)[82] which

[80] Akaike (1973), (1974), (1977), (1985).

[81] The procedure described here is commonly known as "one-variable regression analysis", although Forster does not use the term.

[82] Forster and Sober say simply "sum of squares" (SOS), neglecting to emphasize "corrected". Whereas the "raw" sum of squares equals $\sum y_i^2$ (where y_i is the dependent-variable co-ordinate of the

sums up the squared (vertical) distances from a curve to every data point. It is presumed that the smaller the SOS, the better the fit. Why use the sum of *squared* vertical deviations (SOS), rather than the sum of vertical deviations themselves (SVD) in measuring the degree of fit? The main reason for taking squares is that if the unsquared distances were used, the empirical discrepancies would tend to cancel out and approach zero even for some intuitively ill-fitting curves. But any number of functional transformations other than squares could be used to avoid the cancelling out, such as taking the absolute values, or any even-numbered powers, etc. One should bear in mind therefore that despite SOS-minimisation being the usual criterion of good fit, there is little justification of it in the statistical literature.[83] Those argument that are provided are spurious and at best, depend on a variety of assumptions that are satisfied only in very specific cases (cf. Howson and Urbach, 1993, Chapter 12). As Howson and Urbach show, none of the standard defences of the least-squares criterion works. But there is a Bayesian one that does, namely that

> when the set of possible regressions is restricted to the linear, and one assumes a normal distribution of errors and a uniform prior distribution of probabilities over parameter values, the least squares and the Bayes solutions coincide, in the sense that the most probable value of the regression parameters and their least squares estimates are identical. When the assumption of a uniform distribution is relaxed, the same result follows, though it is reached asymptotically, as the size of the sample increases (Lindley and El-Sayyad, 1968).

Forster and Sober provide no argument for the SOS criterion before (or after) blithely endorsing it. They cannot avail themselves of the Bayesian justification, since they deny that the value of a regression parameter is a random variable with a probability distribution over it. Even if they conceded that it is, the uniform distribution condition would not be justified in the general case, and the case that Forster and Sober, and Akaike before them, are making is quite general.[84] Even more damagingly, the linear-

---

*i*th data point) and is, by itself, quite uninformative, the "corrected" sum of square (CSS) equals $\Sigma(f(x_i)- y_i)^2$ (where $f(x_i)$ is the *predicted* dependent-variable co-ordinate of the *i*th data point, according to the hypothesis $y=f(x)$, and y_i is the *actual* co-ordinate at that point) and measures the goodness of fit of the curve $y=f(x)$ to the observed data. Clearly, Forster and Sober mean the right thing, but their terminology is misleading. Sometimes they speak, more precisely, of the sum of square *deviations* to refer to the same thing. For consistency with direct quotations, I shall use their abbreviation "SOS" throughout.

[83] The most common justification is a well-known theorem of Gauss which connects least-squares estimates to normal distributions. However, this is often taken as a wholesale 'proof' of the least-squares criterion, notwithstanding the very restrictive conditions attached to it by the theorem.

[84] To prevent a possible confusion, two issues must be separated: the distribution of the adjustable parameter (if one treats it, as the Bayesian does, as a random variable), and the distribution of the *x*-

regression assumption that is necessary for defending the SOS criterion will generally not hold, since the AIC (for 'Akaike's Information Criterion') case is supposed to apply to arbitrarily complex regression curves of which the straight line is but a very special case.

The 'true' curve

At least *stating* the SOS criterion (unlike justifying it) is unproblematic. It is more difficult to define what the "true" curve is. In fact, Foster and Sober fail to provide a definition at all. Therefore, I shall attempt to spell it out by looking at the use which they make of the term.

Let us first state what the true curve is *not*. Forster and Sober eschew the 'relativist' option of simply taking the curve supplied by some respectable model, since they want a language-invariant solution that could be used to choose *between* models. But neither can the true curve be the one that fits the data perfectly. To begin with, for each finite data set there would be infinitely many curves "true" in that sense. But further, suppose that a particular curve reflects the true causal law that relates the two variables. Then, since the processes of measurement (of obtaining the data) and, possibly, the measured process itself, are not completely deterministic, the collection of data is bound to include a certain amount of "noise" which causes the data points to deviate from the "true" curve which represents the trend. But can we speak of "the true" curve at all when the process is partially indeterministic, i.e., if there are random deviations from the trend? We can do so only in the sense of the *trend itself*. On its own, this is scant clarification, since the term 'trend' is little clearer than the term 'true curve'.

It would seem natural to define the trend, or the "true" curve as that which traces maximum probability (or, in a case of a continuous distribution, maximum probability density) of potential data points. Each data point will be determined by an equation of the form

---

values of the data points. The former characterises the physical system itself, the way in which its x and y values are correlated. The latter, on the other hand, concerns the way in which the system is manifested to the observer. It is in this second sense that Forster and Sober stipulate that both the "old" and the "new" data sets be drawn from the same distribution, without specifying that distribution's exact form. In other words, the authors, following Akaike, wish to avoid a situation where the observations used to work out a regression are selected through a different algorithm than the observations used to test it, which is fair enough. This, however, should not obscure the fact that in order for the SOS-minimising and the Bayesian solutions to coincide, assumptions must be made about the distribution of the random variable which the authors in question ought to find unbearably restrictive.

$$y = f(x) + \varepsilon,$$

where $f(x)$ is the theoretically predicted value of y (e.g., $f(x)=ax+b$), and ε - the random error which is assumed to be distributed normally with zero mean. As an illustration, consider Galileo's law of free fall as predicting the observed path of an undisturbed body falling under gravity:

$$h = gt^2/2 + v_0t,$$

where g is the constant of acceleration of free fall at the given altitude, t – time, and v_0 – the initial velocity. Not all bodies will have fallen exactly the same distance after the same length of time, even if their starting velocities were identical. However, for every v_0 and t, Galileo's law does give the most probable position of the falling object at every moment (unless the object is very light, e.g. a feather, in which case air resistance becomes a significant force).

And indeed, looking at the way in which Forster and Sober use the term 'true curve', it is clear that they mean the path that a process would take if it were completely deterministic and all observations were error-free. The very first mention of the expression 'true curve' in (Forster and Sober 1994) confirms this interpretation:

> Suppose the true specific curve *determined* the outcomes of the observations we make. Then, if Curve 1 were true, the set of data points we obtain would have to fall on a straight line (i.e., on the straight line depicted by Curve 1 itself) (Forster and Sober 1994, 3).

Later, when discussing the data points determined or 'generated' by the true curve, they also view the true curve as the locus of maximum probability (probability density) points, as described above:

> The most probable outcome [of observing the *Y*-value associated with a given *X*-value – **MM**] is to obtain a *Y*-value that falls exactly on the true curve. Locations that are further off the curve have lower probabilities (symmetrically above and below) of being what we observe) (Forster and Sober 1994, 3).

Based on this implicit definition, what formal properties do "true curves" possess? Because of the random error, the true curve will have a non-zero sum of vertical deviations (SVD) with respect to each finite data set generated by it. However, as the

data set grows, the vertical deviations progressively cancel each other out, their sum converging to zero in the limit. Hence, the *average* value of SVD in every infinite set generated by the true curve is also zero. If the true curve is determined by the function $f(x)$ and passes through (or near) n data points, where the *predicted* co-ordinates of the ith point are $(x_i, f(x_i))$ and its *actual* co-ordinates are (x_i, y_i), then

(1) $$Av(SVD) \equiv 1/n * \sum_{i=1}^{n}[f(x_i) - y_i] = 0$$

In the AIC framework, the average value of SVD is the same as its expected value. Indeed, by rejecting the Bayesian idea of prior probability of data, Forster and Sober effectively assume a uniform prior probability distribution over the data points, which is another questionable and, in the general case, false assumption. Be that as it may, as a result of a uniform distribution, the prior probability of a data point in a sample equals the inverse of the number of data points in that sample: $p(x_i)=1/n$, where $1 \leq i \leq n$. Therefore, the expected value of SVD is simply its unweighted average:

(2) $$E(SVD) \equiv \sum_{i=1}^{n}[f(x_i) - y_i] * p(x_i) = \sum_{i=1}^{n}[f(x_i) - y_i] * 1/n$$

which, for an infinite data set generated by the true curve, is zero.

The introduction of expectations accomplishes two things: it ensures that for each data set, there is *at least* one true curve (by construction, since we are free to tinker with the regression coefficients within $f(x)$ until E(SVD)=0), and that it is *unique* (provided the error term is distributed unimodally, as it is in the normal case to which Forster and Sober largely limit themselves).[85] In fact, the definition of the true curve, implicit in the AIC approach, follows directly from the above assumption that the error term is distributed with zero mean, because the mean itself is nothing but the expected value of the random variable (in this case, the vertical deviation), and the expected value of the sum equals the sum of the expected values. Hence, E(SVD)=0 is entailed by the zero-mean assumption.

[85] Another tacit assumption needed to guarantee uniqueness is one that Forster and Sober actually make in the context of reflecting on 'true curves', namely that we consider only curves of the same polynomial complexity, i.e., only straight lines, or only parabolas, etc. Without this additional assumption, more than one curve could qualify as the true one for a given data set. For instance, take the data points normally distributed around a horizontal straight line. Infinitely many sinusoids would have the same Av(SVD).

The uniqueness (subject to unimodality and the assumption, mentioned above, concerning the functional form) of the curve that satisfies the tacit AIC criterion means that of the infinity of possible curves that connect *all* of the actually observed points, *at most* one is the true curve. On the other hand, the true curve may pass through *none* of the data points and still satisfy the condition. Forster and Sober do make a point of saying that the true curve is *not* that which passes through every possible data point, because the data points incorporate the random error and the true curve does not.

It may seem surprising that Forster and Sober do not spell out the notion of a true curve in this way, i.e. by stipulating condition (1). Instead, they couch their discussion of the properties of true curves in SOS terms, by saying that for each (potentially infinite) data set the true curve minimises Av_{∞}(SOS). Of course, for each data set the minimum E(SOS) converges to a different value, unlike the SVD-based definition where the crucial value is always the same, namely zero. One possible reason for this reluctance may be as follows. In defining goodness of fit, and, as we shall see shortly, in defining distance to the true curve, Forster and Sober use the traditional SOS measure.[86] In general, minimising E(SOS) and setting E(SVD) to zero (which, as I tried to show, is required by the AIC notion of a true curve) will yield different results. There are certain conditions under which minimising E(SOS) yields a curve which traces the points of maximum probability of data, but such conditions are by no means universal.

It is necessary to point out that in the AIC approach, "the truth" is simply that theory which enjoys maximum, albeit not "perfect"[87], correspondence with empirical data (actual and potential), and hence this notion is seemingly quite different from the realist concept of truth which includes reference to states of affairs some of which may be in principle beyond test. However, it is unclear whether it is much weaker than the realist concept, for in the AIC framework, what is true depends on the states of affairs *almost all* of which are observable only in a purely theoretical sense.

It should also be pointed out that in order for condition (1) to be a faithful interpretation of the locution "the true curve", it is necessary that the random deviation from the trend

[86] The answer to this is trivial: due to its mathematical properties, only SVD, but not SOS, makes "perfect" fit with the data possible. The SOS measure always yields a positive number, since each squared distance is non-negative.

[87] It will not be "perfect", on the assumption that empirical fit is measured by the corrected sum of squares (or by any other non-negative measure), since the latter never converges to zero as the amount of data grows - unlike SVD.

be distributed symmetrically around the trend with zero mean. The second clause is satisfied by the definition of the trend, for it is part of the latter that no *systematic* deviations from it are admissible (if the mean of the random error appeared to be $\mu \neq 0$, we would have to conclude that in fact, the *real* trend is μ away from what it appeared to be, while the mean in question *is* zero, after all). But with respect to the first clause, it is questionable whether the symmetric character of the distribution is implied by the definition of randomness or constitutes an independent additional assumption.

The above account of the notion of a true curve required for the AIC approach is an attempt to disambiguate what they actually say. Sometimes they explicitly include the error term in the definitions of particular curves. For instance, speaking of the problem of planetary motion, they say about the interaction of the deterministic law of motion and the error-generating process of observation:

> If the planet's trajectory is a straight line, we can combine these two influences into a single expression:
>
> (LIN) $Y = \alpha_1 - \alpha_2 X - \sigma U$
>
> The last addend represents the influence of error. Here, of course, Y doesn't represent the planet's actual location, but represents its *apparent* location (Forster and Sober 1994, 4).[88]

Strictly speaking, with the error term included, the equation represents not a trajectory but a continuous *family* of trajectories characterised by a probability density distribution (usually it is assumed that the error is distributed normally, with zero mean). Indeed, at each point X_i, the value of Y_i will vary depending on the exact value of σU. In order to represent an individual curve, the random variable U ought to be substituted with a constant. To avoid confusion, I shall distinguish between a *curve* and the *empirical hypothesis* that the curve represents (namely, the hypothesis that actual observations will be distributed, usually normally, around the curve taken to be the true one). The empirical hypothesis, say (LIN), includes the error term that specifies the variance of random deviations from the trend in observed data, but the corresponding curve will not include such a term since it is supposed to represent the trend itself.

[88] In a footnote, Forster and Sober add that if the law of planetary motion itself included a stochastic element, Y would represent the mean position of the planet rather than of our recordings of the planet's position. Thus the authors allow for uncertainty being rooted in the physical system rather than in the measurement process alone.

If all the potential data points had been actually observed, the curve-fitting procedure would be quite straightforward but uninteresting. We must choose the curve so that SOS is minimised. This is no problem, since given enough mathematical complexity, a curve could be constructed that passes through the given points. But since in the "completed" case there is no question of predicting future points, methodologically this case is irrelevant. The only reason we might need to express in this case the relation between the two variables in the form of a function, would be for the purposes of cataloguing. From the scientific point of view, the real interest is in predicting the location of the future data points by means of a function constructed using the points already observed. This is where the curve-fitting problem arises.

When scientists or statisticians try to assess the empirical adequacy of a potentially infinite curve, its degree of fit with the existing data, however it might be measured, is no longer a sufficient measure of its empirical adequacy. For, we cannot guarantee that the best-fitting curve with respect to the existing data will perform just as well as the data set increases to include new cases. Now, in the AIC approach, Av_∞(SOS), the corrected sum of squares averaged over all possible data, is viewed as a more reliable measure of the predictive success of a curve. It is a more reliable measure because the influence of "noise" in the data declines as its amount increases (if it did not decline, we would be dealing with part of the signal itself rather than with noise). This magnitude can be defined (in the AIC spirit, although Forster and Sober themselves do not spell that out), with respect to a particular $f(x)$, thus:

(3) $$Av_\infty(SOS|f(x)) \equiv \lim_{n\to\infty} 1/n \sum_{i=1}^{n} (f(x_i) - y_i)^2 \cdot$$

The averages must be taken, and hence the term $1/n$ must be introduced, to avoid the embarrassing situation in which all infinite curves have the same, namely infinite, SOS. The curve-fitting problem can now be formulated thus: SOS with respect to the actually observed finite set of data is not necessarily a good indicator of Av_∞(SOS), and the minimisation (in the calculus sense) of one is, or can be, quite different from the minimisation of the other. This is, according to Forster and Sober, the predicament of traditional statistics which Akaike's approach is designed to solve:

> The universal refrain is that 'if we proceed just on the basis of the data, we will choose a curve that passes exactly through the data points' [i.e., the curve that minimises the actual SOS - **M.M.**]

whereas

> a curve that is maximally close to the data (because it passes exactly through all the data points) is probably not going to be maximally close to the truth [the truth being defined in terms of minimum expected SOS - M.M.] (Forster and Sober 1994, 6).

Distance from the truth

As we have seen, Av_{∞}(SOS) is intractable and therefore cannot be used *directly* to predict the future empirical adequacy of suggested curves. Before discussing Akaike's proposal on how to circumvent this difficulty, we still need to answer the last remaining question from the above list, namely, how to measure the distance from an arbitrary curve to the true one. Akaike offers the following definition reproduced also by Forster and Sober:

> *Distance from the true curve* T *of curve* C =df
>
> Average [SOS of *C*, relative to data set *D* generated by *T*] -
>
> Average [SOS of *T*, relative to data set *D* generated by *T*][89] (Forster and Sober 1994, 6).

The notion of a data set being "generated by a curve" is not explained by AIC but, as Forster and Sober themselves emphasise, cannot be understood simply as a subset of points on that curve, for then the second term on the right-hand side of the equation would be zero. It cannot be understood as those points that are the most likely with respect to the true curve, because they are on the curve itself and the previous argument applies. But we cannot simply drop "the most likely", because a *continuous infinity* of sets of data points are possible with respect to the true curve and if that were the idea, the notion of *unique* average SOS of *T* and of *C* would be undefined. The only feasible sense in which a data set *D* might be generated by *T* is that *D* is the sub-set of all actually observed data points from a (possibly infinite) set *D'* such that $Av_{\infty D'}$(SOS|*T*) is minimised and $E_{D'}$(SVD|*T*)=0, which means that *T* is the true curve with respect to *D'* and, derivatively, with respect to *D*. But in this case, the locution 'generated by' is misleading because it unduly removes emphasis from the contingent character of *D* and makes one think of it as of something uniquely determined by *T*.

However, this definition faces much graver problems than misleading wording. The definition is based on the following intuition. For every data point, the vertical distance

[89] The last term is greater than zero because of the indeterministic component in the measurement process.

(at that point) of an arbitrary curve from another one (including the "true" curve) analytically equals the difference of the vertical distances from the first curve to the data point, and from the second curve to the data point. This observation applies to distances between two curves *at a particular point.* One could next define the distance *simpliciter* from one curve to another as the average of such point-distances, and it turns out that this distance can be expressed as the difference between the average distances from the two curves to the data set. So far, nothing controversial. But the above observation that holds analytically differs from Akaike's actual definition in one crucial respect: namely, the latter features *squares* of vertical distances in which case, of course, what was previously a hardly helpful but uncontroversial tautology turns into a supposedly illuminating but in reality - highly questionable convention. AIC must defend it on independent grounds, but no such defence is provided (presumably, any such defence would include an argument that the proposed definition of distance between two curves explicates well our intuitive idea of "distance" and is consistent with other widely accepted explications, except possibly with those which AIC are prepared to abolish). Moreover, as I shall argue in Section 3, there are positive reasons to *reject* this definition. Here I shall only point out that given the only plausible meaning of the locution "data set D generated by T", the "distance" from C to T turns out to be non-unique. Indeed, since the exact location of points in the data set includes an element of randomness, for one such D the first term in the right-hand side of the equality may be relatively large while the second - relatively small, and *vice versa*. So, even if there had been no *external* grounds for rejecting the AIC definition (and I hope to show that there are), it would still fail the test of internal consistency.

But let us temporarily accept the above definition for the sake of the argument. It is important to remember that whereas, for the true curve T, E(SVD)=0 (it is defined that way), SOS>0 for any (sufficiently large) data set, and also Av_{∞}(SOS)>0. This difference between SOS and SVD is due to the fact that in the former, random deviations are squared and only then summed, while in the latter, they can take either sign and therefore cancel out. It must also be noted that although Akaike's definition of closeness to the truth deals with actual SOS of finite data sets rather than with Av_{∞}(SOS) of potentially infinite ones, the actual distance is nevertheless intractable since T is never available with any certainty (in fact, the very objective of regression analysis can be defined as finding T, or a good approximation, for a given data set).

But though we can never *calculate* the distance to the true curve from an arbitrary one (and if we could, why, with the true curve at hand, would we want to?), Akaike purportedly has found a way of *estimating* it. In fact, although for purposes of prediction

AIC always envisage the use of individual curves, they consider the principal units of epistemic analysis to be *families* of curves - all with the same number of adjustable parameters[90] but differing with respect to values of the attached coefficients (so that e.g. all quadratic functions, in the AIC view, belong to the same family). Choosing a family constitutes, according to AIC, the first step in the standard curve-fitting process.[91]

Once we have settled on a particular family, we can select a unique vector of coefficients that minimises SOS (at least, it will be unique in sufficiently low-dimensional families, and those are the only ones that scientists usually consider). This is the second and final step. Its rationale is that within each family of curves characterised by the same number of adjustable parameters (and, we should add, by the same functional form), the curve that minimises SOS is assumed by AIC to also minimise Av_{∞}(SOS).[92] In turn, Av_{∞}(SOS) is an explication of our idea of the degree of empirical adequacy, future as well as present, of a curve. If mathematical complexity is no object, the *actual* SOS *can* be reduced to zero (or brought within an arbitrarily small distance from zero) by increasing the number of adjustable parameters, i.e., by selecting more and more complex families.

However, this would have very undesirable implications to our ability to predict *future* data points. If the number of adjustable parameters is too large (i.e., the family of curves chosen is too complex), Av_{∞}(SOS) increases even as (actual) SOS decreases. As an example, consider the following picture. Initially we are given 5 data points, with x ranging from 0.15 to 1.8. Using the simplest (linear) regression model, we could represent the data points with a straight line (using the best-fitting straight line). It is clear, though, that the fit of the regression line with the data points is far from perfect, as the predicted $f(x)$-values differ quite significantly from the actual y-values. One might be tempted to use a cubic

---

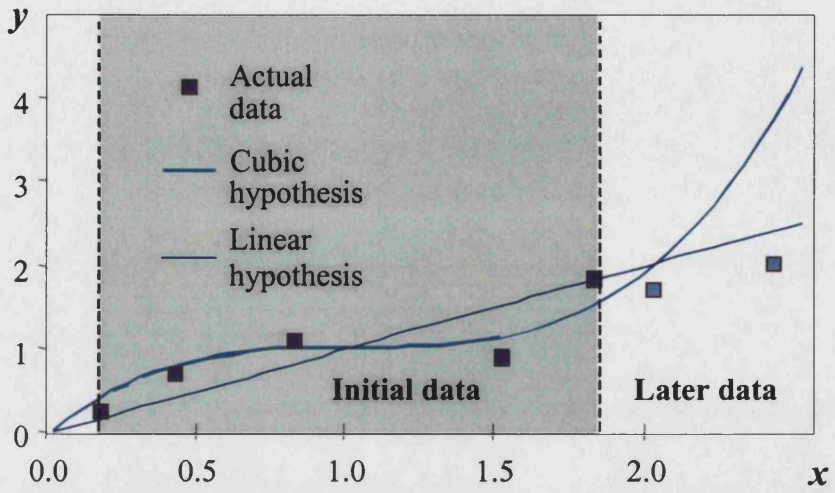
[90] One must also add (which Forster and Sober do not, at least not explicitly) to the condition of equality of the number of adjustable parameters also the condition of identity of functional form. Otherwise, the equations $y=ax+b$ and $y=a\sin x^b$ would seem to define the same family, as each contains two adjustable parameters.

[91] In fact, in most of the literature no deliberation is apparent as to which regression model (in AIC terms, which family of curves) to choose. Rather, statisticians start off with the linear assumption and work with it unless the fit that it affords proves to be too poor, in which case a slightly more complex model is selected. Jeffreys made a similar point in his (1961), pp.100-101. Of course, Forster and Sober scould argue (although they do not explicitly do so) that the default status of the linear model is itself a result of the preliminary, imprecise assessment in which fit is weighed against simplicity. It is this weighing that is purportedly explicated by Akaike's theorem, to be stated shortly.

[92] It could be objected that a very complex curve could have a zero SOS with respect to a finite data set but a large Av_{∞}(SOS). Such a possibility is precluded by selecting the family (and therefore, the degree of complexity) of the curve first.

regression model, as the best-fitting cubic curve affords much better fit with the actually observed data than the best straight line.

Figure 4: Linear vs. cubic approximation of sample data



However, as we obtain data from outside the original range, i.e., $x>1.8$, the upward-pointing branch of the cubic function deviates farther and farther from the new data points, whereas the straight line, its fit remaining quite imperfect, nevertheless stays close to them. If the trend in the distribution of new data does not change, the cubic curve will deviate from the data more and more. Hence,

$$Av_{\alpha}(\mathrm{SOS}|f(x)=x^3-3x^2+3x) > Av_{\alpha}(\mathrm{SOS}|f(x)=x),$$

while

$$\mathrm{SOS}(f(x)=x^3-3x^2+3x) < \mathrm{SOS}(f(x)=x)$$

with respect to $0>x\geq 1.8$.[93]

This phenomenon is referred to by statisticians as "overfitting" and, according to Forster and Sober, illustrates the more general notion of *ad hoc*-ness in that a temporary gain in empirical adequacy is achieved at the price of a dramatic decline in predictive accuracy.

[93] Of course, the example was deliberately constructed so that outside the initial sample range, the linear hypothesis engendered better fit than the cubic one. The opposite picture could have been devised just as easily.

For the obvious reason that we cannot know the exact locations of future data points, Av_{α}(SOS) cannot be directly estimated and therefore cannot be used to assess the predictive value of a particular curve. (Indeed, even if we were given the "true" function that determines y_i for each i, the exact value of Av_{α}(SOS) would still be intractable. Above, following AIC, "truth" is defined (in this limited context) as that curve (or the function which generates that curve) whose expected value of the sum of vertical distances from the data is zero - in effect, as "open-ended" empirical adequacy "on average". Thus, the "true" curve only shows the most probable[94], not the actual, position of each data point and cannot even be used to calculate SOS with respect to a finite sub-set of data - we have to refer to the data themselves). This definition links the notion of truth very closely to that of predictive success and thus differs from that of the realist, but it is the notion that Akaike works with. Also, this definition allows AIC to use Akaike's results based thereon to measure the same predictive properties of curves that would have been most naturally measured by Av_{α}(SOS) had that quantity been tractable.

The phenomenon of over-fitting was known to statisticians long before Akaike; Akaike's purported achievement is that he formally explained it. First, it must be pointed out that each family includes *two* linear terms: a linear occurrence of the independent variable, and of the "error" variable. The latter results from the partially indeterministic character of the data-generating process.

Overfitting, in the AIC framework, is the result of a mistaken attempt to incorporate the noise into the trend (we do assume that there is noise - i.e., that either the process being studied is not completely deterministic, or measurement is subject to random error, or both). Overfitting occurs when the error term is treated as another regression coefficient (which, by assumption, it is not),[95] as a result of which the curve adjusts itself more to the noise in the data and less - to the trend. Hence, a sure sign of having missed the "true" (or, better, "trend") curve (with respect to which E(SVD)=0), is getting a very small value for SOS in a sufficiently large subset of the data. Just how small is "too small", it is claimed, can be estimated from Akaike's theorem reproduced below.

---

[94] More precisely, it shows the position of maximum probability density, since point probabilities as such are everywhere zero.

[95] Technically, the error term can also be viewed as an adjustable parameter (indeed, Forster and Sober do so). However, unlike the 'proper' adjustable parameters, i.e. regression coefficients, its expected value is zero. The error term does have non-zero variance.

Previously, distance from the truth has been defined for individual curves. The distance from the truth of a *family* of curves can then be defined, following AIC, as the average of the distances from the truth of the best-fitting (for each particular sub-set of data, as the number of data points increases) members of that family. This finally allows us to formulate Akaike's result:

> Estimated (Distance from the truth of family F) = SOS $[L(F)] + 2ks^2$ + Constant.
>
> $L(F)$ is the member of the family that fits the data best[96], k is the number of adjustable parameters that the family contains, and s^2 is the variance (degree of spread) of the distribution of errors around the true curve. The last term on the right hand side is common to all families, and so it drops out in comparative judgments (Forster and Sober 1994, 9).

This result (reproduced by Forster and Sober without proof) features two important characteristics of any family of curves: its goodness of fit, and its degree of complexity. Akaike's theorem states that the verisimilitude of a theory is determined as a trade-off between its *actual* empirical adequacy and its complexity. It is implied by Akaike's theorem that overfitting hurts precisely because the error term is multiplied unnecessarily: in order to capture all the fine details of the trend, the theory amplifies the noise, so that the gain in actual empirical adequacy is more than off-set by the increase in future noise. To sum up, Akaike allegedly explained why, other things (mainly, observed empirical adequacy) being equal, "simpler, more unified and less *ad hoc* theories" (those with fewer adjustable parameters) are preferable.

One, however, would be ill-advised to accept this result too uncritically. A number of disturbing questions remain unanswered:

(1) What is an "estimate" of the distance between two curves (or between a curve and a family)? Remember that in the above definition of distance (contentious as it is), we are given only the meaning of *actual* distances. Does "estimated" simply mean "the most probable"? Or does it mean, as seems likely, "expected" (i.e., average, with respect to a particular prior probability distribution over all possible distances)? Neither of these interpretations is supported by Forster and Sober's texts (moreover, they seem to be inconsistent with their general methodology that eschews the use of prior probabilities), nor is any alternative interpretation hinted at.

[96] In the sense that it minimises the SOS with respect to the actual data.

(2) How are we supposed to calculate "the variance of the distribution of errors around the true curve", given that the true curve itself is not given? And we do need the numerical value of this variance in order to make comparisons between different-dimensional curves where the lower-dimensional one enjoys better fit to the data.

THE BEARING ON BAYESIANISM

In Akaike's approach, as Forster points out, we are moved towards the *truth* - by maximising expected empirical adequacy. Distance from the truth, as defined above, can be minimised by finding well-fitting curves that are not too complex. Opposed to it is the central tenet of Bayesianism, as Forster sees it - namely, that we are only concerned with the *probability* of truth rather than with actually moving *towards* it. As Forster and Sober put it,

> The fundamental principle behind Akaike's method is that we should aim to select hypotheses that have the greatest predictive accuracy. Since the truth has the maximum possible predictive accuracy and accuracy is a measure of 'closeness', Akaike's recipe aims to move us towards the truth. In contrast, the central thesis of the kind of Bayesianism we will criticize here is that hypotheses should be compared as to their probability of truth (Forster and Sober 1994, 22).

But are the differences as big as Forster and Sober make them out to be? I shall argue that they are more perceived than real, and that the charge that Bayesian conditionalisation fails to take us closer to the truth is an unjust one. It anything, it is the AIC approach that downplays the truth as the ultimate epistemic desideratum. As Sober writes in his (1999),

> The Akaike framework assumes that inference has a specific goal; the goal is not to decide which hypothesis is most probably true, or most likely, but to decide which will be most predictively accurate' (Sober 1999, 5).

How does that constitute 'moving towards the truth'? Greater predictive accuracy is not necessarily greater truth. In fact, Forster and Sober view a gap between the truth and predictive accuracy as more than just a theoretical possibility, notwithstanding the above statement in their (1994). This is evident from another excerpt from the same article of Sober's:

> The truth can be a misleading predictor. It is a familiar fact that idealizations are valuable in science when a fully realistic model is either unavailable or is mathematically intractable. The Akaike framework reveals an additional virtue that idealizations can have – *even when we possess a fully realistic (true) model, a (false) idealization can be a better predictor*' (Sober 1999, 6 – italics added).

Truth vs. probability

On the positive side, there are good reasons to believe that striving towards greater probability does lead us closer to the truth. I shall argue that the putative reasons that Forster and Sober produce for the alleged impossibility of this do not hold water.

Why not conditionalise?

First, I shall very briefly re-state the traditional Bayesian approach in statistical reasoning which is the object of Forster's attack. Suppose we are trying to estimate the value of a particular parameter. It can be either a numerical coefficient in a highly theoretical formula, or an observable quantity (I do not hereby endorse a rigid "theoretical-observational" distinction but use the terms as they are normally used, i.e., contextually). We begin by assigning to the different possible values of the parameter prior probabilities (or, in the continuous case, probability densities). Then we conduct an observation (or an experiment) while defining the likelihoods of the alternative hypotheses about the value of the parameter, with respect to the different outcomes of the observation (experiment):

$$\text{likelihood}\ (H_i|E_j) = \text{probability}\ (E_j|H_i),$$

where H_i is a hypothesis stating that the value of the parameter is the *i*th element of the partition of the range of possible values of the parameter, and E_j is the *j*th outcome of the observation (experiment). After the actual (say, *m*th) outcome has been observed, the new ("posterior") probability (or probability density) of each hypothesis is defined according to Bayes Theorem and the Rule of Conditionalisation:

$$p(H_i) = p(H_i|E_m) = p(H_i)*p(E_m|H_i)/p(E_m),$$

where $p(E_m)$, or the "prior probability of evidence", equals, according to Bayes theorem of probability calculus, to the prior probability-weighted sum of the likelihoods that the

evidence receives from all the alternative hypotheses.[97] According to a number of so-called "convergence theorems", as we iterate this procedure, the upward-bulging curves representing our successive probability (or probability density) functions will grow steeper, with the point of global maximum shifting closer to the true value of the parameter (in the limit, centring on it), *regardless* of the shape of the original probability (or probability density) function. This procedure is essentially what Forster has in mind when he criticises Bayesian statistics in the light of Akaike's results (we shall see in a moment how he does that and on what grounds).

It can be noted that the contrast between Akaikeism and Bayesianism is not as sharp as Forster and Sober present it. When the Bayesian strives towards greater certainty, he is thereby moving closer to the "true" theory and distancing himself from the false ones. Indeed, epistemic probability can plausibly be interpreted as a negative measure of the distance separating the knowing subject from a full acceptance of a theory. Moreover, the convergence theorems mentioned above show that, if the true theory is already among the set of alternatives between which we are trying to adjudicate, then as the amount of data grows, Bayesian conditionalisation will lead to the probability of the true theory converging to unity, no matter what the original probability distribution over the set of alternatives was. (More will be said about the Bayesian notion of verisimilitude in the last section.)

For Forster's and Sober's opposition between Bayesianism and the theory that they advocate to be valid, there must be something wrong with the Bayesian using *probabilities* as milestones in his arduous pursuit of the true theory. Indeed, in Forster's view the very Bayesian procedure described above is suspect.

Prior probability and logical entailment

It is an unalterable fact, say Forster and Sober, that subject to some minimally acceptable degree of empirical fit, scientists prefer simpler theories. Presumably, Bayesians would like to be able to explain this fact. There are two sources from each the Bayesian derives the final probability assignment: prior probability, and likelihood. Could the preference for simpler theories be reflected in the likelihood? No, say Forster and Sober, since the likelihood can always be increased by choosing *more complex* theories. If the preference for simpler theories be expressed at all, it can only be done

[97] In the continuous case, we take the *integral* of the prior probability density-weighted likelihood function over the entire range of the possible values of the parameter.

through the prior probabilities of theories, so that a simpler theory receives higher prior probability which may or may not be off-set by its lower likelihood.

According to Forster and Sober, this avenue is also closed: if the Bayesians were to account for the role of simplicity as an epistemic desideratum by rewarding simpler theories with higher prior probability, they would run into a logical contradiction. In the crucial paragraph in this regard they write:

> The key element of any Bayesian approach is the use of Bayes's Theorem, which says that the probability of any hypothesis H given any data is proportional to its prior probability times its likelihood: $p(H|\text{Data}) \propto p(H)*p(\text{Data}|H)$. However, it is an unalterable fact about probabilities that (PAR) is more probable than (LIN)[98], relative to any data you care to describe. No matter what the likelihoods are, there is no assignment of priors consistent with probability theory that can alter the fact that $p(\text{PAR}|\text{Data}) \geq p(\text{LIN}|\text{Data})$. The reason is that (LIN) is a special case of (PAR). How, then, can Bayesians explain the fact that scientists sometimes prefer (LIN) over (PAR)? (Forster and Sober 1994, 22)

A similar argument is made in Forster (1995). In that article Forster considers the following equations each determining a family of trajectories: 'H_1: $ds/dt=0$; H_2: $d^2s/dt^2=0$; and H_3: $d^2s/dt^2=g$, where g is an adjustable parameter, and s is the distance of free fall of some projectile' (in time t)[99]. He then states that, since the solutions to these three equations contain one, two and three adjustable parameters, respectively, H_1 is simpler than H_2 which in turn is simpler than H_3.

Since H_1 is simpler than H_2 (provided that we accept the number of adjustable parameters as the criterion of simplicity), it must be more probable, but according to the probability calculus, it cannot be, since H_1 *logically* entails H_2. This entailment is supposed to hold because for every arbitrary solution of H_1: $s=a$, there is a pair of values of the adjustable parameters in the solution of H_2, namely $b=0$ and $c=a$, that turns it into the solution of H_1. 'This is very embarrassing to Bayesians', rejoices Forster.[100]

---

[98] These are the hypotheses that the true curve is a parabola or a straight line, respectively - **MM**.

[99] Forster (1995), p.407.

[100] This reasoning is meant to discredit, among other Bayesian strategies, Harold Jeffreys' 'simplicity postulate' more on which later.

However, the rejoicing is premature. The trouble with these arguments is that, contrary to appearances, the entailment relations that Forster and Sober postulate between (LIN) and (PAR) and between H_1 and H_2, simply do not hold. This is very easy to show by looking at the logical structure of, say, the first pair:

(LIN): $\exists a{\neq}0\ \exists b\ \forall x\ (f(x){=}ax{+}b)$

(PAR): $\exists a{\neq}0\ \exists b\ \exists c\ \forall x\ (f(x){=}ax^2{+}bx{+}c)$

These are two existentially-quantified statements neither of which logically implies the other (this could be shown by constructing a tree with open branches in it). *Pace* Forster and Sober, (LIN)$\not\subset$(PAR), thus rendering their argument unsound.[101]

I shall return a little later to the reason why it may erroneously appear that (LIN)$\subset$(PAR), and argue that Forster trades on the ambiguity of the expression "scientists prefer" as applied to theories. There are at least two senses in which scientists might prefer less simple theories to their logically stronger reformulations: they may deem them either (1) more probable; or (2) more useful (in a sense to be explained later). Of course, scientists *never* "prefer-1" logically stronger theories, nor should they: a proposition cannot be more probable than one of its consequences. But this does not mean that scientists may not "prefer-2" if, in their eyes, lower probability is outweighed by other virtues, say, greater informativeness. It is this second sense of preference which is relevant here.

I shall say more about the notion of "preference-2" (which is closely related to Maher's[102] concepts of "acceptance" and "cognitive utility") after looking at how Forster himself sees the options left by his analysis to the Bayesian. Forster considers two possible Bayesian defence strategies and argues that they are untenable. One of them is based on assigning probabilities to *families* of curves, the other - to *individual* curves. I

---

[101] Cf. Howson (1973). One could argue that the condition $a{\neq}0$ in the definitions of (LIN) and (PAR) is illegitimate, and that without it, the first hypothesis does entail the second. I argue in the next sub-section why the condition is indeed necessary (although Forster and Sober do not themselves state it). Here, I shall limit myself to pointing out that we are concerned not with pure geometry, but with scientific methodology. Accepting the $a{=}0$ case in the definition of (LIN) would amount to saying that, contrary to appearances, the 'dependent' variable does not, in fact, depend on the independent variable. Then why are we bothering to fit the curve at all? And once we have agreed to use the non-degenerate functional form of the simplest curve, i.e. (LIN), we should do likewise for each subsequent type of curve.

[102] Maher (1993)

shall begin with the former and show that the two strategies are, in fact, but two different phrasings of a single Bayesian response.

First of all, let us note that the argument that probability is a diminishing function of simplicity, is not due to Forster. It was Popper (e.g. 1959, 381) who first developed it in response to the proposal of Jeffreys and Wrinch (1921), and the invalidity of his reasoning is argued e.g. in Howson (1973), Howson (1988), and Howson and Urbach (1989, 292). They show that Popper's argument hinges on some gratuitous assumptions about the nature of probability, such as the thesis that the more testable a theory, the less probable it is. There is, however, a difference between Popper's and Forster's arguments. The former does not specify in what logical relation the more and the less simple theories whose probabilities are being compared stand, and hence imports the extra assumptions criticised by Howson and Urbach and by others. Forster, on the other hand, explicitly limits his case to situations where the more complex theory is implied by the simpler one, and therefore rests the argument on probability calculus alone. My contention, however, is that the limitation of scope leaves the argument altogether inapplicable.

What are we comparing?

Let us now look specifically at the case of (LIN) vs. (PAR). Are they the proper theoretical alternatives? Whereas (LIN) is indeed a sub-family of (PAR) and thus automatically possesses a lower or equal probability, it is not (PAR) itself that is the interesting and relevant subject of comparison with (LIN), but rather the complement of the latter (PAR*)=(PAR)\(LIN), and since there is no entailment relation between (PAR*) and (LIN), the problem disappears: we are constrained in assigning prior probabilities to these two subfamilies only by the probability calculus. The same holds for H_1 versus H_2. Any solution of H_1 is of the form $s=bt+a$, where $b\neq 0$ and a is determined by the initial conditions, while its legitimate rival H_2* is solved also by $s=bt+a$, where $b=0$ and *ditto* for a. Since H_1 does not imply H_2*, we are free to reflect the former's greater simplicity in a higher prior probability.

Forster finds this proposal unacceptable for three reasons. Firstly, he says that by comparing (LIN) with (PAR*) rather than with (PAR), we do not solve the original problem:

> In response, we note that this *ad hoc* maneuvre does not address the problem of comparing (LIN) versus (PAR), but *merely changes the subject* (Forster and Sober (1994), 23).

Secondly and thirdly, Forster doubts that Bayesians can justify an assignment of probabilities to families obtained in such a way, and that they can define likelihoods for families of curves.

In my view, these objections do not hold water. If the Bayesian 'changes the subject', it is because the subject needs changing. By focusing our comparison on (PAR*) rather than on (PAR), we are comparing two subclasses of a universal class rather than one such subclass with the universal class itself, which seems both intuitively rational and represents the usual practice.[103] To begin with, the Bayesian may compare whatever theories he likes. Each of the theories mentioned above has its probability, and he simply records them. Having done that, the Bayesian selects those of them that he deems, in the given theoretical context, the appropriate subjects of comparison. That (LIN) and (PAR*) are, indeed, appropriate subjects of comparison, and that the shift to this comparison from the one advocated by Forster and Sober, namely between (LIN) and (PAR), is far from *ad hoc*, is clear from the mathematical form of the corresponding hypotheses. I have made my point with respect to H_1 and H_2*. A similar point can be made with respect to the other pair of families of hypotheses. The mathematical form of the hypothesis (PAR) is

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \sigma U,$$

while that of (LIN) is

$$Y = \alpha_0 + \alpha_1 X + \sigma U,^{104}$$

which results from the parabolic equation by setting $\alpha_2=0$, while (PAR*) results from the same equation by setting $\alpha_2\neq 0$. Suppose that we have accepted (PAR) as true or

---

[103] To use a very down-to-earth example, suppose we are facing a choice between different varieties of the quarter-pound hamburger. Forster urges us to choose between Burger King's Whopper and a quarter-pounder 'in general', while Bayesians would rather choose between a Whopper and *another* quarter-pounder, say, Mac Donald's Big Mac.

[104] In both cases, σ^2 is the variance of random deviations from the curve. It is worth noting that *pace* Forster, both equations above determine, for each pair of α_0 and α_1, a *family* of curves rather than an individual curve. The reason is the presence of the random term σU which will result in data points being off the curve determined by the "deterministic" part of the equation.

very probable. We might still find it insufficiently informative and try to supplement it by a content-increasing addition of a hypothesis about the value of α_2. Surely, two hypotheses differing as to the value of α_2 are perfectly appropriate objects of comparison. In fact, testing two hypotheses such that one stipulates that a particular value is zero while the other - that it is not, against each other, is the most common type of hypothesis-testing in classical statistics. Also, it is unclear what is wrong with *ad hoc*-ness to begin with. True, Forster and Sober make a point that *ad hoc*-ness in the sense of overfitting to past data is detrimental to predictive success. However, this sense of the terms clearly does not apply in the present situation since there are no adjustable parameters involved.

From the Bayesian (and more generally, from any probabilistic) perspective, the fact that the probability of the universal class is not greater than that of any of the elements of its partition, is not very interesting: the statement "p(LIN)$\leq p$(PAR)" is a tautology, and therefore we would not even be asking whether or not it is true. What we *will* be asking is whether or not the statement "p(LIN|PAR)$\leq p$(non-LIN$_i$|PAR)", for some i, is true (including the case of (non-LIN$_i$) which coincides with (PAR*), the complement of (LIN) with respect to the general parabolic hypothesis).

I should point out though that such a procedure need not be available only to the *subjective* Bayesian. As mentioned in the previous chapter, it has been argued (notably by Popper) that the prior *logical* probability of a parameter being *exactly* equal to any particular value (including 0) is zero. But the measure-theoretic considerations due to which the probability of α_2=0, apply only if the probability distribution of α_2 was a Lebesgue measure.[105] However, there is absolutely no reason in general to choose a Lebesgue measure. Furthermore, even if we did have a reason, in a particular context, to think of α_2 as distributed in a way rendering its probability zero, we could still assign to it a non-zero probability density, which is sufficient for an application of Bayesian procedures.

As for the task of *justifying* particular probability assignments to (LIN) and (PAR*), further on Forster recognises that it is not incumbent on the subjectivist Bayesian to provide such justifications. Furthermore, I shall argue below that some forms of justification are available also to the subjectivist.

[105] In fact, Popper equated logical probability with the Lebesgue measure.

Forster and Sober's final objection to the employment of Bayesian procedures in curve fitting seemingly has more bite. It is not enough to assign prior probabilities in a consistent way; likelihoods must also be defined in a way that is consistent both internally and with the assignment of priors. Again, Forster and Sober employ the 'curve *vs.* family' dichotomy. The Bayesian should either explain how likelihood could be defined for a family of curves, or restate the whole procedure (including the assignment of priors) in terms of individual curves. With this latter approach, in their view,

> The trouble is that a *particular* curve, as opposed to a family of curves, cannot be assigned a value of k [where k is the number of adjustable parameters - MM.] *on a priori grounds*. After all, any curve is a member of *many families* of different dimensions[106].

Purportedly, this argument compels the Bayesian to devise the whole conditionalisation procedure in terms of families rather than individual curves, and Forster and Sober think that they already know how to discredit that alternative. However, in my view their argument accomplishes nothing of the sort. Indeed, Forster and Sober's inference is invalid. Their second statement is undeniably true if trivial, but it doesn't entail the first one, namely the alleged impossibility of assigning a k-value to an individual curve. Since an individual curve has no free variables, k=0 for each one.

Once it is established that individual curves are identical with respect to their k-values, some may be tempted to argue that each curve must therefore receive identical prior probability, which would render 'factoring simplicity into the priors' impossible. I shall now turn to showing that such a suggestion would be groundless.

"Informationless" priors

It is a widely held view that, in the absence of reasons to do otherwise, the Bayesian should adopt 'informationless', or uniform, prior probability distributions. Of course, this is nothing other than the Principle of Insufficient Reason clad in newer clothes. In Bayesian literature, this view was expressed, among others, by Harold Jeffreys in his *Theory of Probability* where, incidentally, he also introduced the idea of measuring simplicity by the number of adjustable parameters.[107]

---

[106] Forster and Sober (1994), 25

[107] "The complexity of a law is now merely the number of adjustable parameters in it, and this number is recognizable at once..." (Jeffreys, 1961, 100).

Jeffreys was concerned with applying Bayesian procedures to the task of testing a law whose functional form includes one or more adjustable parameters.[108] This involves assigning prior probability to each theoretical alternative, which presents two sorts of difficulties. The first one consists in assigning prior probability to each *functional form* of the law based on the number of adjustable parameters involved, which Forster and Sober contend is an impossible task. The second problem is deciding, for each functional form, the likeliest *value* of each parameter. In the former context, Jeffreys put forward his Simplicity Postulate, more of which later. First, I shall discuss Jeffreys' take on the problem of estimating the true value of a parameter, once the functional form has been fixed (albeit provisionally).

In discussing the best rule for assigning prior probability to each value of a parameter, Jeffreys considers three cases, depending on the range of the parameter in question. The cases are: from a to b (a finite interval of the real line); from $-\infty$ to $+\infty$ (all real numbers); and from 0 to $+\infty$ (all non-negative numbers). The first two cases can be treated as one, due to the symmetry of the metric properties of the end-points which is lacking in the third case:

> If the parameter may have any value in a finite range, or from $-\infty$ to $+\infty$, its prior probability should be taken as uniformly distributed. If it arises in such a way that it may conceivably have any value from 0 to $+\infty$, the prior probability of its logarithm should be taken as uniformly distributed. [...] The essential function of these rules is to provide a formal way of expressing ignorance of the value of the parameter over the range permitted. They make no statement of how frequently that parameter, or other analogous parameters, occur within different ranges. Their function is simply to give formal rules, as impersonal as possible, that will enable the theory to begin (Jeffreys, 1961, 102).

In other words, the purpose of a uniform (or log-uniform) prior distribution is to start the process of Bayesian conditionalisation in as unbiased a way as possible. But is the assumption of a uniform distribution not just as biased as any other? Does it not constitute a substantive judgement about the value of the parameter? Jeffreys thinks not:

> The answer is really clear enough when it is recognized that a probability is merely a number associated with a degree of reasonable confidence and has no purpose except to give it a formal

[108] For a long time, Jeffreys considered only differential equations with rational coefficients, the mainstay of 19th-century physics, to be the 'proper' form of a scientific law. However, he later realised that this was too restrictive.

> expression. If we have no information relevant to the actual value of a parameter, the probability must be chosen so as to express the fact that we have none. It must say nothing about the value of the parameter, except the bare fact that it may possibly, by its very nature, be restricted to lie within certain definite limits (Jeffreys, 1961, 102).

Thus, according to Jeffreys a uniform prior distribution is merely a statement of the fact that we equally lack certainty in relation to each possible value (or range of values). However, applying a uniform distribution in the infinite and semi-infinite cases involves assigning infinitesimal probability to every finite interval. In order to avoid dealing with ratios of infinitesimals when comparing the probability of a variable falling into different finite intervals, in infinite and semi-infinite cases he proposes to represent certainty with $+\infty$:

> There is no difficulty in this because the number assigned to certainty is conventional. It is usually convenient to take 1, but there is nothing to say that it always is. But if we take [in the semi-finite case - **M.M.**] any finite value of σ, say α, the number for the probability that $\sigma<\alpha$ will be finite, and the number for $\sigma>\alpha$ will be infinite. Thus the rule would say that whatever finite value α we may choose, if we introduce Convention 3[109], the probability that $\sigma<\alpha$ is 0. This is inconsistent with the statement that we know nothing about σ.
>
> This is, I think, the essence of the difficulty about the uniform assignment in problems of estimation. It cannot be applied to a parameter with a semi-infinite range of possible values (Jeffreys, 1961, 103-104).

For this reason, in the semi-infinite case Jeffreys proposes to take logs. The log function is convenient in that its domain is from 0 to $+\infty$, which is perfectly suited for the semi-infinite case, and its range is from $-\infty$ to $+\infty$, thus avoiding the above problem and making probability metrically identical in the infinite and semi-infinite cases.

Having taken care of the additional problems presented by the semi-infinite case, Jeffreys acknowledges the difficulties traditionally associated with uniform distributions in general. The main one is that non-linear functional transformations of uniform distributions are not themselves uniform:

---

[109] In fact, Jeffreys' Convention 3 ("If p entails q, then $P(q \mid p)=1$" - Jeffreys, 1961, 21) has nothing to do with it. The result follows from the definition of conditional probability: $P(\sigma<\alpha \mid 0<\sigma<\alpha) = P(\sigma<\alpha \ \& \ 0<\sigma<\alpha) / P(0<\sigma<\alpha) = x/\infty = 0$ for any finite x. $P(0<\sigma<\alpha) = \infty$ and x is finite by stipulation.

> ...if a parameter is unknown then any power of it is unknown; but if such a parameter is v, then if v lies between v_1 and v_1+dv, we should have according to the rule
>
> $$P(v_1 < v < v_1+dv \mid H) \propto dv,$$
>
> and if we try to apply the rule also to v^n we should say also
>
> $$P\{v_1^n < v^n < (v_1+dv)^n \mid H\} \propto dv^n \propto v_1^{n-1}dv$$
>
> The propositions considered on the left are equivalent, but the assessments on the right differ by the variable factor v_1^{n-1} (Jeffreys, 1961, 104).

Among the real-life examples of where this problem has arisen, Jeffreys mentions measuring the charge of an electron: "Some methods of measuring the charge of an electron give e, others e^2, but de and de^2 are not proportional" (*ibidem*).

Therefore, instead of making, with Bayes and Laplace, the prior probability distribution of v proportional to dv, Jeffreys proposes to make it proportional to dv/v. Its advantage is that for every n, dv/v and dv^n/v^n are proportional with a constant coefficient. Indeed,

$$dv^n/v^n = nv^{n-1}dv/v^n = ndv/v.$$

The constant coefficient n is then absorbed into the coefficient of proportionality in the definition of the probability measure. Analogously, log-uniform distributions on semi-infinite domains are invariant with respect to power transformations:

$$dv^n/(v^n \log v^n) = nv^{n-1}dv/(nv^n \log v) = dv/v \log v.$$

On this measure, the distance between any two points does not depend on the power used. But the proposed solution is not sufficiently general. It is intended for power-functions of the variable being estimated; however, for the vast majority of monotone functional transformation this will not work. For instance, if x is distributed uniformly, $\log x$ is not; there is no general reason why we might not need to measure logs of a certain parameter. Therefore, Jeffreys' suggestion merely limits the scope of the problem but does not eliminate it altogether. The lack of invariance under monotone transformation, which was the bane of the Principle of Indifference in its initial formulation, remains.

But this is not the only objection against the uniform distribution principle. Why consider only three possible cases: a finite range, $-\infty$ to $+\infty$, and 0 to $+\infty$? It is quite conceivable that our prior knowledge of a parameter implies a range say from -1 to $+\infty$, or from $-\infty$ to 1, or some other semi-infinite range that includes both positive and negative values. This renders both the uniform and the log-uniform rules inapplicable. Or else, in certain classes of problems the conceivable range of the adjustable parameter may be the union of a finite and a semi-infinite interval, say, $(-1;0)U(1;+\infty)$. How is one to apply, simultaneously, both the uniform and the log-uniform measure to the same domain? Jeffreys himself was aware of the limited applicability of the uniform/log-uniform rule, and realised that its application presupposed very restrictive notions of scientific inquiry:

> I think that at this point we come up against one of the imperfections of the human mind that have given trouble in the theory: that it has an imperfect memory. If everything that attracted its attention was either remembered clearly or completely forgotten it would be much easier to make a formal theory correspond closely to what the mind actually does, and therefore there would be less need for one. Data completely forgotten would then be totally ignored, and we know how to do that [through the principle of uniform distribution – **M.M.**]; those perfectly remembered could be used in the theory in the usual way. But the mind retains great numbers of vague memories and inferences based on data that have themselves been forgotten, and it is impossible to bring them into a formal theory because they are not sufficiently clearly stated (Jeffreys, 1961, 107).

In other words, we may often be inclined to view certain values in the range as more probable than others, but cannot quite articulate why. Jeffreys suggests that we ignore such unclear preferences, but this seems unnecessarily restrictive and does not achieve anything. For one thing, as explained above, severe conceptual difficulties remain with accepting the principle of uniform distribution as a universal standard of rationality (even subject to Jeffreys' provisos, such as that we must have no clear reason to view the distribution as non-uniform). Moreover, in practice the applicability of such a principle would be severely limited anyway. Finally, no one prevents a scientist from adopting a uniform distribution in the context of a given inquiry should he so wish; but equally, there are no convincing reasons to force him to adopt such a distribution.

A further objection against uniform distribution as an overriding principle is that, as we shall see, it is inconsistent with Jeffreys' own Simplicity Postulate (and, more damagingly, with any finite probability assignment to an arbitrary theory with bound variables).

Any of these difficulties make it impossible to adopt the rule of uniform priors as a universal methodological principle. This leaves the Bayesian free to assign prior probability to alternative curves howsoever he sees fit. Nevertheless, I shall argue that should the Bayesian wish to incorporate the *k*-value into his prior probability assignment, he would still have good reasons for choosing, in a non-trivial manner, the most "representative" family for each curve, allowing him to apply the Simplicity Rule if and when the context of inquiry warrants such an application.

A trouble with likelihood?

Let us first restate the predicament which Forster attempts to show the Bayesian to be in: families of curves cannot have well-defined likelihoods (because for different probability distributions over the values of adjustable parameters, we purportedly get different likelihood values for each given data set - cf. the "language-dependence argument" below), while individual curves cannot have rationally assigned priors (because each curve belongs to a number of families of different dimensions).

Forster[110] rejects the attempts to define the likelihood of a family as a weighted average of the likelihoods of its member curves. The weight would have to be the prior probability densities of individual curves. He considers two cases: where such a density distribution is uniform, or "informationless", and where it is not.

> If p(Curve|F) is strictly informationless, then it is easy to see that p(Data|F)=0. Almost every curve in the family will be very far from the data. This means that if we accord equal weight to every curve in F, the average likelihood of F will be zero[111].

This conclusion is a clear *non sequitur*. Although almost every curve in the family will indeed be "very far from the data", each will be only *finitely* far. Hence, every individual likelihood, and thus the average likelihood (no matter how it is weighted) will be finite, although for those probability densities which do not peak close to the maximum-likelihood curve, the latter will be very low indeed (it is partly for this reason that we would often seek some prior information about the density distribution for individual curves). Therefore, there is no need to entertain, as Forster and Sober do, the proposal to limit the calculation to the "area of non-zero likelihood". Further, Forster

[110] Forster (1995), p 408-410

[111] Forster and Sober (1994), 23

and Sober claim that a *non*-uniform prior density distribution would not allow us to define a unique family likelihood. Their argument proceeds from an alleged lack of invariance under reparametrisation. Let us critically examine this claim:

> Consider the following pair of equations:
>
> (LIN) $Y = \alpha_0 + \alpha_1 X + \sigma U$
>
> (LIN') $Y = (\alpha_0')/3 + (\alpha_1')/2X + \sigma U$
>
> These equations define exactly the same family of straight lines. Yet, the proposal [to define the likelihood of a family as a weighted average of the likelihoods of its constituent curves -**M.M.**] entails that the latter has 6 times the average likelihood of the former.[112]

That this striking result also follows from fallacious reasoning, is clear from Forster's and Sober's own footnote. We are asked to integrate

> a function $f(x)$, where $f(x) = 1$ between the limits 0 and 1, and $f(x) = 0$ elsewhere. Clearly,
>
> $$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$
>
> Yet if we transform coordinates such that $x'=6x$, while equating $g(x')$ and $f(x)$ for corresponding values of x and x', we obtain
>
> $$\int_{-\infty}^{\infty} g(x')\mathrm{d}x' = 6$$
>
> (Forster and Sober 1994, 24).

Let us recall that both the independent and the dependent variables in regression analysis are always expressed in *units*, be those seconds and metres or unemployment and crime % rates. The gist of the procedure described above is redefining our units of measurement of the input. Inevitably the units of measurement of the *output* must be re-scaled accordingly. E.g., if electricity costs one £10 a month, one shouldn't be surprised to spend £60 for a half-year. The monthly cost of electricity does not change whether we count consumption monthly or semi-annually. Analogously, as long as the density distribution over particular curves within a family does not depend on the notational form in which they are described (and there is no reason why it should), the prior-

[112]Forster and Sober (1994), 24

probability weighted average of individual likelihoods provides a unique measure of the likelihood of the family, *once the prior distribution has been fixed.*[113]

In fact, Forster (1995) seems to recognise that only a *non-linear* transformation of the probability (or density) function (and not merely its notational re-description) could possibly support his case for language-variance. If we choose a prior distribution so that the maximum likelihood curve receives a fairly high prior weighting, the likelihood of the family will be greater than if we had chosen a distribution that gave the maximum likelihood curve a low prior weighting. But, as I have argued, no further language variance occurs once the prior distribution has been fixed. Granted, as a consequence, we cannot ask about the likelihood of a family *before* introducing a prior distribution on it. But nor should we: before saying how likely the evidence is in light of a particular family (or "model", as is more customary to say in regression literature), it is necessary to specify first how that model is structured. After all, before deciding how likely a football team handicapped by *n* players is to concede a goal within the next *m* minutes, we would like to know, among other things, what proportion of remaining players will be playing in defence, in midfield, etc (i.e., we would be inquiring about the team's structure). Closer to the issue at hand, it is universally conceded even by non-Bayesians that the *prior* probability of evidence is a function of the prior probability (or density) distribution over the theory space. There is no principled reason why the *posterior* probability of evidence (likelihood) could not be, in certain cases, similarly influenced. Therefore, the assumptions which make it is mathematically possible to define the likelihood of a family as a weighted average of the likelihoods of individual curves, by integrating the probability density function, are by no means unusual or unreasonable ones.

Next, I would like to point out that Forster himself, following Akaike, defines the notions such as "the distance from the truth" and "predictive power", traditionally associated with fully parameterised theories, for *families* of curves (theories) - by equating them with the corresponding values of the best-fitting *members* of such families (or, when there are different best-fitting members with respect to different data sets, by taking *unweighted* averages). One could therefore argue that, once this approach

[113] Curiously, while Forster and Sober make so much of the alleged lack of 'language-invariance' of Bayesian approaches to simplicity, it has recently been pointed out the AIC solution itself faces the equally grave problem of lacking in 'world-invariance'. Take two hypotheses, 'All emeralds are green', and 'All emeralds are grue'. If 'green' is taken to be the primitive predicate, the first hypothesis turns out to be simpler than the second, but the result is reversed if 'grue' is taken is primitive and 'green' is defined in its terms (cf. DeVito 1997).

has been taken with respect to the notions of verisimilitude and predictive power, there is nothing further to be lost by extending it to the notion of likelihood.[114] I do not advocate defining the likelihood of a family as that of its best-fitting member, although it would solve the problem of weighting: all the weight would be assigned to *one* particular curve. This procedure would involve no reference to a particular prior probability or probability density distribution, and would therefore destroy Forster's objection. However, this option in effect amounts to setting the probability (or density) of the maximum-likelihood value of the parameter to one, and of those of all the other values thereof - to zero. If one deems this consequence unacceptable, as a Bayesian would, then this particular response to Forster's purported difficulty collapses, but not before the basic notions of the AIC approach on which Forster's attack was based in the first place.

The Principle of Simplicity

Now it appears that Forster's and Sober's skepticism about the Bayesian's ability to 'factor simplicity into the priors' is unwarranted. However, if we did wish to reward simpler theories, other things being equal, with higher prior probability, how would we go about it? The most common way to do that is by means of Jeffreys' above-mentioned Simplicity Postulate. First, I would like to argue that individual curves, as well as their families, are legitimate entities to which the Simplicity Postulate may be applied.

Curves or families?

From the above, it is clear that Forster's misgivings about the family-based Bayesian strategy are unfounded. But his scepticism about the individual curve-based strategy is equally unjustified. Granted, each curve belongs to an infinite number of families of different dimensions; the n-th family results from equating the coefficient attached to the highest-power term in the equation of the n+1st family to zero. Among all such families, there is the *simplest* one - such that the curve in question is obtained by substituting non-zero numbers for each of the coefficients in that family's equation. For a straight line, such family is (LIN), and so on. That a higher-dimensional family could be obtained by introducing 'virtual' non-linear terms with zero coefficients attached to them, is of no consequence since the newly-introduced terms bear no information. The following information may be useful. Just as it takes not more and not less than two

[114] It is even more puzzling in light of Forster's remark on the difficulties of assigning prior probabilities to individual curves where he equates the (prior) probability of a curve with that of the family (-ies) of which it is a member.

separate points to define a straight line and three – to define a plane, it takes exactly two adjustable parameters to define the "family" of linear regression models, etc. Granted, sometimes one *can* define a straight line by *more* than two points - provided that all the extra ones lie between the first two. But one does not *need* the extra points. And anyway, in such a case, the points in question can no longer be used to define a plane or a higher-dimensional hyperplane. They have zero information content relating to the expansion of the straight line into a higher dimension (analogously to the "virtual" non-linear terms in the equation of (LIN) having zero coefficients); for the same reason that they are consistent with the form of a straight line, they are useless in defining a plane etc. Claiming that a curve belongs to the family that has the same dimension as itself *in the same sense* in which and *to the same extent* to which it belongs to the infinite number of higher-order families, is absurd. It is absurd because it amounts to ignoring an extremely relevant piece of information - that the "virtual" terms are all zero. It is the essence of the Bayesian approach that none of the relevant information must be disregarded, and even the opponents of Bayesianism dare not attack this principle. Hence, as long as one accepts that "free" (in epistemic as well as in economic cost) relevant information is always welcome, there is no way of denying that for each curve there is at most one family whose prior probability should be used as a guide to the prior probability of the curve itself.

To sum up, just as we can unequivocally assign prior probabilities to families (possibly taking into consideration their degrees of complexity in relation to the shape of the data), we can do so for individual curves, as well. So, whether the Bayesian chooses to operate with families or with individual curves as his primary concept, there are meaningful and consistent ways of defining both prior probabilities and likelihoods, and therefore, of applying the customary Bayesian techniques also to the curve-fitting problem.

Why the 'principle'?

It is now clear that, given two rival theories, the scientist may well define one of them as the simpler alternative, and in the absence of any other salient information, give it the higher probability and proceed with conditionalisation from there. But does the ability to form *pairwise* rankings entail an ability or even obligation to form complete probability orderings of *all* possible theories based on their simplicity, howsoever it may be defined? Let us examine Jeffreys' motivation behind his affirmative answer to this question.

Previously, I discussed Jeffreys' views on estimating the value of an adjustable parameter within a given functional form of a putative law.[115] But regardless of how we distribute probability between all the possible values of each of the adjustable parameters, we must probability-rank the functional forms themselves before we can calculate the prior probability of each specific law containing no free *variables*. Jeffreys' Simplicity Postulate is a proposal to make the prior probability of a functional form inversely related to the number of adjustable parameters in it. The rationale behind this is to make the total probability of all putative laws pertaining to a particular domain finite, which requires the probability of open law-like sentences to be a convergent series. Suppose, says Jeffreys, we failed to adopt what I shall call the 'decreasing probability' rule.[116] Bearing in mind that the total probability must remain finite, each functional form would have to receive infinitesimal probability:

> If all laws had the same prior probability it would be infinitesimal, and would remain infinitesimal on any amount of evidence. Thus there could be no stop, not even a temporary pause, unless we agree that every law has a finite prior probability. But then if there are an infinite number of possible laws their prior probabilities must form a convergent series (Jeffreys, 1961, 100).

If the series did not converge, total probability would be infinite, in violation of the probability calculus.

> This result implies the possibility [rather, necessity - **M.M.**] of arranging possible laws in an order of decreasing prior probability. What can this order be? [...] It is the order in which the laws ordinarily arise for consideration, that of increasing number of adjustable parameters. This principle of convergence was what Wrinch and I originally called the simplicity postulate (Jeffreys, 1961, 100).

Let us look more closely at the relation between the Principle of Uniform Distribution of parameter values, discussed previously, and the Principle of Simplicity. Initially, it

---

[115] Jeffreys himself speaks simply of 'laws'. However, I find this misleading, as the term 'law' refers properly to a sentence without free variables, with each adjustable parameter fixed at a certain value. This does not contradict the fact that law-like statements with adjustable parameters are *bona fide*, truth-valued sentences, provided every variable is bound by a quantifier.

[116] Clearly, if the probability of a functional form were *directly* related to the number of adjustable parameters, the series would diverge, and even infinitely small probabilities of each member would result in infinitely large total probability. Therefore, Jeffreys does not even consider this case, concentrating instead on the uniform distribution over the number of adjustable parameters.

may seem that the Uniformity Principle, although proposed and defended by Jeffreys on independent grounds, is a logical consequence of the Principle of Simplicity. Indeed, the latter stipulates that any two laws with the same number of adjustable parameters shall have the same prior probability, and since two laws differing only in relation to the value of an adjustable parameter do have the same number of adjustable parameters, they should receive the same probability. On the other hand, imagine that the only two alternatives meriting serious consideration are (LIN) and (PAR*). Suppose (LIN), being simpler, receives probability 2/3, while the remaining 1/3 goes to (PAR*).[117] We are interested in the prior probability distribution over the value of the coefficient in (PAR*)'s quadratic term. Since a priori the coefficient may range from $-\infty$ to $+\infty$, each value must receive the same probability. If we allow only rational coefficients, each possible value will have infinitesimal probability, including 0. But the hypothesis (LIN), and therefore the value 0 of the quadratic coefficient, has already received probability 2/3. Thus, Simplicity and Uniformity are actually inconsistent.

However, things are even worse than that. As the above example illustrates, the Uniformity Principle entails that if the total probability of the set of parabolas is finite, then the total probability of the set of straight lines is infinitesimal. In general, for any two embedded theories of complexities n and n+1, if $p(T^{n+1})$ is finite, then $p(T^n)$ is infinitesimal. Inductively, for a theory to receive finite probability, it must have infinitely many adjustable parameters, which is absurd. Understood as a universal principle of 'logical probability', Uniformity must go, for this reason and for others discussed previously, although nothing prevents us from using it as a contingent assumption in limited contexts.

In fact, it is possible to argue that Jeffreys himself did not intend for his probability distribution rules, Uniformity and Simplicity, to apply across the board. From Jeffreys' statements quoted above it is clear that in order for a prior probability ordering on theories to be possible (in more than the trivial sense where each one receives zero probability), there must be at most countably many possible parameter values. Thence Jeffreys' and Winch's initial restriction of putative laws to differential equations with rational coefficients, of which there are indeed countably many. Once we drop this unsustainable restriction, uncountable infinity rears its ugly head, rendering the probability of each theory zero. As Colin Howson points out, Jeffreys found, although did not always express clearly enough, an alternative way of restricting the number of possible parameters to at most denumerable:

[117] In fact, *any* split of total probability between (LIN) and (PAR*) will do.

> How does one find an at most countable set which contains all the possible laws governing the data which science will ever deem worth testing, and which "will excluded no possible law *a priori*"? So put, the question has a rather obvious answer, whose explicit acceptance characterises the difference between Jeffreys's earlier and later varieties of Bayesianism. This answer is that the possible laws science will ever deem worth testing are those which it actually will test, not those which populate some independently-given logical space, the vast majority of which will never even be thought of.
>
> Simple though it is, this observation clearly provides the solution Jeffreys sought to the cardinality problem, since no more than a denumerable sequence of laws could, even through infinite time, ever be actually tested. On the other hand, the human mind is in principle capable of considering any intellectual structure whatever as a hypothetical explanation of the observational data it is confronted with, and so the condition that no law is excluded *a priori* is plausibly satisfied; for there is no limit to the number of possible laws from whose ranks a candidate for serious consideration may at any time emerge (Howson 1987, 74)

Once it is decided that it is only the actually entertained theories to which the Principle of Simplicity is supposed to apply, in any particular context we can decide, *a priori*, how many adjustable parameters would be 'too many' in any plausible theory of the subject matter, and thus safely rule out those theories that are considered too complex. Total probability can then be easily distributed among the finite number of the remaining theories, giving each one finite probability.

Up to now, I have argued that, *pace* Forster, the Principle of Simplicity that Forster traces to Jeffreys is not logically or methodologically inconsistent, so long as its scope is limited to the set of theories put forward in the context of a specific inquiry. However, having rejected Jeffreys' own arguments for the Principle of Simplicity, I feel it necessary to at least indicate the lines along which a *positive* Bayesian justification of the said principle (suitably restricted in scope) might be given. Let us consider the epistemic motivations (other than the fear to run foul of Akaike's theorem) that might prompt the Bayesian to "factor simplicity into the priors". I shall present one "pragmatic" and one "historicist" justification (although there may well be, and probably are, more than one of each kind).

A "pragmatic" justification

On a "sufficiently small" domain, the graph of a complicated function can be very closely approximated by that of a simpler one; in the limit, for the two-dimensional case, by a straight line. Hence, on such domains there is no expected empirical loss from using simpler curves: they are "just as good". On the other hand, with each new adjustable parameter in the equation of the curve the probability of making a mistake increases (although does not necessarily *strictly* increase), assuming that each adjustable parameter is estimated independently.[118] Since most scientists would rather not expose themselves to a greater risk of getting it wrong when the extra risk is not compensated for by significantly higher accuracy, they tend to favour simpler theories, other things being equal. Such epistemic preference, in turn, translates into higher subjective probability, as described in connection with Savage's theory.

The cynic would say that this argument is merely a crude and non-technical restatement of the Akaike argument which is also based on a trade-off between empirical accuracy and complexity in determining the prior probability of theories. The cynic would be largely right, with two qualifications. The first one is that in my version the 'loss function' associated with each additional adjustable parameter is not decreasing closeness to the truth of the augmented theory, as in Akaike, but its decreasing probability. Forster and Sober are themselves at pains to emphasise the difference between greater closeness to the truth and greater probability; while the former is eminently attainable, the latter, in their opinion, is not. The alleged reason for this is that for each of the 'nested' theories the probability of being the true one is exactly the same as for the others, namely zero. This view, in turn, hinges on the same assumptions that beset Jeffreys' initial writings on simplicity and which, Howson argues, he eventually came to reject: namely, the identification of probability with logical probability, and the related application of probability measures to unrestricted domains.

The second difference from Akaike is clear from the above; namely, the prior probability that depends on the scientist's reluctance to commit himself to more risky estimations than strictly necessary is the overtly subjective one. AIC, on the contrary, couch their argument in deceptively objectivist terms.

[118] Even if the parameters are not estimated independently, the extra probability of error in estimating more parameters is still non-negative, although possibly smaller than in the case of independent estimation.

Not that the subjectivist approach needs, or indeed can, eschew any mention of objective logical constraints. On the contrary, as Ramsey emphasised, subjective probability subsumes deductive logic. The subjectivist fully agrees with Forster and Sober that (LIN) is less probable than (PAR), because the former entails the latter. At the same time, on the account developed here, (LIN) is *a priori* more probable than (PAR*), since it has fewer parameters (each of which bears the possibility of an estimation error). One could argue that, because in the equation of (LIN) there are infinitely many constant terms (all, except possibly one, set to zero), my argument would render the prior probability of (LIN) zero and therefore not greater than that of (PAR*). This, however, would be a rash conclusion since all the constant terms can be treated as a vector constant (indeed, this is how such "multiple" constants are treated in mathematical analysis - regardless of whether or not the constants m and n in a differential equation equal zero, they are "lumped together" into a single constant k, and for all intents and purposes the equation is treated as having just one constant term). With this in mind, in the second pair of families the matters are straightforward. As we have seen, probability theory produces a definitive judgement about the relative prior probabilities of (LIN) and (PAR*) while deductive logic has nothing to say about the matter since the two theories are contraries (or, if we take (PAR) to be the universal class, contradictories) of each other. But in the first pair, there seems to be a clash between these two lines of reasoning. While the relation of logical consequence (or of class inclusion) entails that (PAR) is more probable, it nevertheless seems in need of a "penalty" for having more adjustable parameters than (LIN). However, this clash is merely apparent. The Bayesian argument from the number of adjustable parameters has been epistemic rather than logical - i.e., based on the consequences of *estimating* a parameter rather than of simply having it in the formula. But of course, since, so long as we are dealing with (PAR) as such rather than with (LIN) or (PAR*), the value of a_2 *is not* being estimated (otherwise we would already be at the level of sub-families, contrary to the assumption). Therefore, (PAR) is not epistemically disadvantaged compared to (LIN) and hence nothing stands in the way of deductive logic-based considerations.

A "historicist" justification[119]

Even if scientists once began their enterprise without subscribing to something like the Principle of Simplicity, in other words, they did not let the considerations of simplicity affect their prior probability functions, they must have abandoned that neutrality by

[119] Elliott Sober has drawn my attention to the fact that a similar historicist justification of simplicity has been proposed by Nozick in his essay 'Simplicity as Fall Out'.

conditionalising on the results of previous research. Historically, many or most of the major successful (in the sense of being accepted by the scientific community as a paradigm for future research) law-like generalisations in the natural and social sciences turned out to be quite simple. For instance, the dispersion of many types of energy in 3-D spaces obeys the cubic law; consumption (in the Keynesian model) is a linear function of income; etc. In other words, in the past it has paid off to construct theories as simple as possible, provided the chosen level of simplicity guarantees sufficient empirical fit. Surely scientists ought to take that into consideration when trying to repeat past successes.

Of course, these two lines of reasoning, the 'pragmatic' and the 'historicist', are intended as proving not that the Bayesian *must* assign higher prior probability to simpler theories "come what may", but that he *can* do so in specific contexts.[120]

[120] In private correspondence Elliott Sober has objected that mere possibility is not enough: 'I agree that Bayesianism 'can' assign priors to reflect simplicity considerations in such a way that intuitive results ensue. But what Bayesianism needs to do is to show which priors *should* be assigned. Otherwise it isn't really providing an account of the role of simplicity' [my italics – **M.M.**]. This is a very common criticism of Bayesianism, and one to which Ramsey and more recently Howson and Urbach (1993) responded already by likening probability theory to deductive logic. Deductive logic provides the rules of deductive inference, untroubled by the fact that it has nothing to say about the truth values of the premises. Analogously, probability theory provides the rules of inductive inference, and it would be asking too much to require it to provide the probability values of *its* premises.

BEYOND AKAIKE

The problem allegedly facing the Bayesian in the light of Akaike's results, has been that simplicity, no longer a semi-aesthetic desideratum but a determinant of expected empirical accuracy, ought to be given due weight in any successful scientific methodology, while Bayesianism is, according to Forster, unable to do so due to the problems outlined in Section 2. I have tried to argue that those problems are fictional. We *can*, if we like[121], reward simpler theories with higher prior probability without either assigning greater probability to logically stronger theories or leaving our basic concepts undefined.

How simple is simplicity?

Still, it is worth explaining why, whereas the role of simplicity as an epistemic desideratum is immediately *entailed* by Akaike's theorem, it is "merely" strongly supported by the Bayesian methodology. I shall now argue that this fact results from a more general (involving fewer extra-methodological presuppositions or commitments) character of the latter. In fact, Sober acknowledges that AIC may be limited in both theoretical validity and in practical application:

> However, we have to admit that our argument is only as strong as its premises, and not every statistician will agree that AIC stands on firm foundations (Forster and Sober 1999, 7).

The premises that the authors refer to concern the technical underpinnings of Akaike's theorem.[122] But they also concede that on the methodological level, 'one still might wish for more' than the AIC improvement on Fisher's views on statistical inference (*ibid.*). I would like to emphasise that this latter admission concerns not the technical assumptions of Akaike's theorem but the methodological construction built by Forster and Sober on its basis. My contention is that, *pace* Forster and Sober, the further improvement on Fisher that they refer to is already available, in the shape of suitably

[121] Moreover, in most circumstances, it would be irrational not to do so, as I argued at the end of the previous section.

[122] To simplify *in extremis*, these premises are (1) the uniformity of the underlying distribution of the data set, and (2) the normality of the distribution of the random variable. Whereas (1) may be viewed as a condition of the very existence of the 'true' curve, and therefore of the possibility of a regression, (2) enjoys no such 'transcendental' status and is genuinely restrictive. In order for Forster's and Sober's methodological constructions to get off the ground, the underlying technical assumptions must be much more general than that.

refined Bayesianism. In the course of this argument, I shall demonstrate how the notions of informative content and of closeness to the truth can be developed from a decision-theoretic Bayesian perspective.

That the AIC approach may be based on even more limited foundations than its authors are prepared to admit, is suggested by this approach's straightforward equation of simplicity with the number of adjustable parameters. In such a case, the hypotheses (LIN) and (SIN) must be equally simple, since each one is determined by an equation with just two adjustable parameters:

$$\text{(LIN)} \qquad Y = \alpha_0 + \alpha_1 X + \sigma U$$

and

$$\text{(SIN)} \qquad Y = \alpha_0 + \alpha_1 \sin X + \sigma U,$$

respectively[123]. However, something tells us that the linear hypothesis is simpler, at least in the conventional sense, because it has fewer "kinks" (namely, none). Besides, in science sinusoidal processes are generally considered more complex than linear ones, by virtue of having such non-linear characteristics as amplitude and phase. This indicates that the main intuition on which the AIC approach is based, may be flawed.[124] As Jeffreys hinted, the link between functional form and simplicity is only a tenuous one:

> ...it appears much better not to restrict the possible types of law at all, but merely to be ready for them as they may arise for consideration, whatever their form (Jeffreys, 1961, 100).

I do not wish to deny that the number of adjustable parameters reflects some of the essential intuitions about the degree of simplicity of a scientific law. My contention is only that as an explication of the idea of simplicity as it features in scientific discourse, this one is neither perfect nor complete.

---

[123] As pointed out by Donald Gillies, one could represent $\sin X$ as a power series, which would give (SIN) a more complicated form than (LIN). However, a power series expansion, despite consisting of infinitely many terms, does not contain any new adjustable parameters, hence the suggestion is of no use.

[124] To be fair, Akaike and Forster-Sober are far from unique in measuring simplicity by the number of adjustable parameters; this seems to be a fairly common view. Jeffreys, for one, held it, although in a less rigorous form, e.g. he viewed the linear function and the parabolic one as equally simple despite the number of adjustable parameters being one and two, respectively.

Cognitive utility

So far, I have argued why, from the Bayesian point of view, scientists often prefer-1 (LIN) to (PAR*) (i.e., deem the former more probable). Now, as promised, I turn to the reasons why scientists sometimes prefer-2 (LIN) to (PAR) *despite* a lower probability of the former. In so doing, I shall rely on the notion of acceptance of a theory developed recently in a number of Bayesian writings, such as Maher (1993) and Festa (1996).

The former, citing a number of case studies from the history of science, argues that high probability is neither necessary nor sufficient for the acceptance of a theory. As we remember, Forster's contention has been that if probability were used to measure relative merits of competing theories, scientists choosing more contentious theories over logically weaker ones, would be irrational. It would indeed be so if Bayesians believed that probability is all there is to rational acceptance. But they do not believe so, because otherwise they would require that scientists (and all the rest) accept only informationless tautologies. Indeed, for every powerful hypothesis with a posterior probability *close* to one, there is a tautology whose probability is *exactly* one and in whose favour the powerful hypothesis would have to be rejected.[125] What they do believe is that probability is *one*[126] of the crucial factors playing a role in rational acceptance. If scientists' cognitive decisions (which theory to accept and which to reject, and how to modify those theories that seem promising but not quite adequate in their current form, etc.) obey a number of *prima facie* plausible assumptions (such as connectedness and transitivity), then several *representation theorems* can be proven that show that in their cognitive decisions, scientists maximise their expected cognitive utility with respect to some probability function that is unique, and to some cognitive utility function that is unique up to a positive affine transformation. That, consequently, each act of choosing one theory over another must have a numerical utility associated with it, should not seem controversial to anybody who subscribes to the Bayesian decision theory in its general form and further, treats theory acceptance as a (cognitive) decision. Naturally, we have to say something about the nature of cognitive utility.

A useful outline of a Bayesian treatment of cognitive utility is given by Patric Maher in his *Betting on Theories* (1993). According to Maher, who on this point agrees with Popper, scientists like high informative content in their theories. Without being a

[125] E.g., 'All ravens are black' *vs.* 'All black ravens are black'.

[126] It becomes the only factor when some background assumptions are satisfied, such as in the case of choosing from a number of equally informative hypotheses.

substantialist, one can simply say that the intellectual enterprise we know as "science" came about as, and continues to be, a quest for true (or at least "empirically adequate") information. Hence, informativeness cannot help being valued by scientists. But of course, Tolkien's trilogy ("Lord of the Rings"), being far more informative than most historical treatises I know of, is nevertheless not considered to be an exemplary work of a historian - exactly because it is not true nor empirically adequate (one could argue what the exact relation is between truth and empirical adequacy, but it is reasonable to suppose that a fair degree of empirical adequacy "on average" is at least a *necessary* condition for a theory being true). We can take that into consideration, following Maher, by defining cognitive utility separately for such acts as "accepting theory T if it is true" and "accepting theory T if it is false". Maher insists that cognitive utility for the acts of the first kind be an increasing function of the informativeness of T, assuming that scientists aim at as full truth as possible, in other words, at finding informative true statements rather than at collecting true but useless tautologies. As for those of the second kind it is reasonable, but not necessary, that cognitive utility be non-increasing (and possibly, even strictly decreasing: most of us, including scientists, feel the more tricked by accepting a lie the more detailed it is) [127]. Apart from these fairly modest stipulations, we cannot say much about the exact shapes of the cognitive utility functions that scientist might have. As an example, Maher cites the following family of cognitive utility functions for a theory A:

$$u_k(A, r) = kc(A)\text{-}d_r(A),$$

where $c(A)$ is the informative content of A and $d_r(A)$ - its distance from the truth, given that the true state of the world is r. Finally, k "represents the relative weight this utility function puts on the desideratum of informativeness, as compared with the competing desideratum of avoidance of error" (Maher 1993, 144).

This example shows, among other things, that *pace* Forster, the Bayesian has a non-*ad hoc* way of taking due account of the importance of *closeness to the truth* as well as of the probability of being true. (Even if there is no reliable way of ascertaining how close to the truth we are, in other words, even if "closeness to the truth" is epistemologically inaccessible, it is still cognitively relevant - counterfactually, a "normal" scientist would

[127] If we want the cognitive utility of false hypotheses to depend also on how false they are, i.e., on their verisimilitude, then we get a (possibly continuous) set of cases of the form "accepting theory T if the real world is in the state X". Unlike the "true/false" case where all that mattered was whether or not $X \in \boldsymbol{T}$, where $\boldsymbol{T}$ is the set of states consistent with the theory T, the scientist now has to have preferences for as many cases as there are discernible Xs. Maher proves a representation theorem for the continuous case as well.

prefer to hold a false theory that is comparatively close to the whole truth than another false theory that is comparatively far from is). Especially (but not exclusively) in the case of a statistical hypothesis, the probability of hitting on the true (or "trend") hypothesis is vanishingly small, so it makes sense to define cognitive utility also for those curves that are closer to or farther from the true one, and Maher's approach does that.

Now we can define expected cognitive utility just as we define usual expected utility, namely, as a probability-weighted sum (or, in the continuous case, the probability density-weighted integral) of (cognitive) utilities. As a result, a rational (=expected cognitive utility-maximising) scientist will accept (LIN), or in other words, prefer-2 it to (PAR), unless his k, or the weight he places on informativeness, is unusually small.

CHAPTER 4: MARKET BEHAVIOUR

The case for Bayesianism being the preferred *scientific* methodology, can be approached by examining to what degree it is followed by persons who are universally considered rational, in *non-scientific* contexts. Such an examination is usually carried out by creating fairly simple experimental situations and checking whether the subjects make their choices in accordance with the axioms of Bayesian decision theory (such as transitivity and separability). A failure to do so in a substantial number of cases is taken to indicate that the axioms in question, and therefore expected-utility maximisation (which follows from those axioms by the Representation Theorem), are not requirements of rational thought and action.

Ramsey and De Finetti already realised that simple experimental arrangements in a psychologists lab will have methodological import only as long as the subjects are kept genuinely interested in the outcome. In other words, the differences in utility associated with the outcomes of the subjects' choices, must be perceptible. This is clearly not the case with most of the attempts to undermine transitivity etc. It has been argued more recently (Cosmides 1989) that even simple rules of deductive logic, like *Modus Ponens* and *Modus Tollens* (whose relevance to rationality has never been questioned) will be consistently violated if the subjects are disinterested in the outcome of the experiment. Therefore it is important to look at those cases where the opportunity cost of making the wrong decision is high, and few areas of decision-making penalise mistakes more than playing on the stock market. While an individual may do badly on the stock market just through sheer bad luck, it is safe to assume that at the aggregate level contingent circumstances play a much reduced role.

"FAIR PRICE"

Two main explanations of the behaviour of shares deserve, due to their historical prominence and practical importance, special consideration: that share pricing exhibits an underlying trend, and that such pricing is completely random. According to the Efficient Market Hypothesis (EMH)[128], agents form expectations of the likely performance of

[128] Cf. Samuelson (1965), Fama (1970), (1976), (1991), LeRoy (1989).

quoted companies over a particular time horizon, and "present-discount" the expected streams of profits (and therefore of dividends[129]) generated by that performance. Such present-discounted values of expected dividend streams constitute what agents perceive as the "objectively warranted" or "fair" price of the share, *given the information available.*[130] We can see that by assuming that the fair price of a share is the price at which we expect it to sell at a later date, plus the dividends we expect it to yield, all discounted back to present. Using "p^*_t" and "p_{t+1}" to denote the fair price and the actual price of a share at times t and t+1 respectively, "d_{t+1}" for the dividend payable at the latter time, and "Φ_t" for the information available at t,

$$p^*_t = (1+\rho)^{-1} (Ep_{t+1}+Ed_{t+1}|\Phi_t),$$

where $(1+\rho)^{-1}$ is the present-discounting factor with ρ usually interpreted as the nominal (i.e., uncorrected for inflation) interest rate.[131] If we assume, as the EMH does, that actual prices either randomly oscillate around their objectively warranted values (or even equal those values at all times, as in the martingale model to be introduced shortly), and hence, use the fact that $E(p_{t+1}) = p^*_{t+1}$, we can iteratively substitute away p_{t+1}, p_{t+2}, etc, to get present fair value in terms purely of present-discounted expected future dividends:

$$p^*_t = \Sigma(1+\rho)^{-i} Ed_{t+i}.$$

The "objectively warranted" values of shares govern the seemingly chaotic behaviour of the stock market (and can therefore be also called "fundamental values" of shares). Since in this model agents form rational, probability-based, expectations of the potential returns on investing in each particular share, and since they amend their expectations, according to the rule of conditional probability, as new information becomes available, one would be justified in calling it the Bayesian Model of the stock market (BM).

---

[129] It is immaterial that varying percentages of company profits may be paid out as dividends: as we shall see shortly, what matters is not dividends as such, but dividends *plus capital appreciation.*

[130] If the time horizon is finite, the (present-discounted) resale value of the share will enter into calculations.

[131] We use the nominal rather than the real (i.e., net of inflation) interest rate in order to incorporate both the positive real gain of holding the share, and the prevented real loss we would have suffered from inflation had we held cash instead of buying the share.

Opposed to EMH is the Random Walk Hypothesis (RWH) according to which share prices fluctuate, to a high degree, randomly and are not based on any "fundamental" value. Any endogenously non-random element in share price behaviour is due entirely to the quirks of mass psychology.[132]

It is possible to argue that one of the factors that informed this anti-essentialism about stock performance was Keynes's idea of "animal spirits" whose ebb and flow is a dominating influence on stock market behaviour. While it would be a stretch to characterise Keynes as an early proponent of RWH, a shared sentiment is clearly discernible.

However, it is quite possible that, had Keynes lived at a different time, he would have been less impressed by 'animal spirits' and other imponderables. Indeed, the 1930s, when most of his economic research and practical investing were done, was an exceptional decade from the point of view of capital markets and the underlying economic performance.[133] As was first reported by Officer (1973), the volatility of the Dow-Jones Industrial Average (a volume-weighted average of twelve largest stocks traded on the New York Stock Exchange) between 1930 and 1939 was about three times as much as in 1897-1929 and in 1940-1969 (as measured by the standard deviation of monthly returns). A very similar disparity was observed in the volatility of US industrial production in those periods. European, especially British, indicators would not have been very different. All this suggests that the economic data on which Keynes based his views of capital markets, were very unrepresentative.

According to Pratten (1993), the jury is still out on the question of empirical adequacy of EMH *vs.* RWH. This is not surprising, since given the complexity of the subject matter, rival economic models tend to be difficult to adjudicate. But what interests the *methodologist* (as opposed to the economist), is the underlying model of agents' reasoning employed by EMH and RWH. Even if there are sufficient market imperfections to render

[132] I shall postpone a more formal presentation of RWH until after discussing the notion of martingales.

[133] Previously, 20th-century economic history had usually been divided into "pre-war" and "post-war". Officer's distinction seems to be much more accurate.

Market Behaviour

EMH a less then perfect account of the functioning of the stock market, it may well be the case that agents (individual and institutional investors) form their expectations of the "objectively warranted" share prices which they use in their investment decisions, and that they conditionalise such expectations on new information.

There are numerous reasons why an agent might want to buy a stock perceived by him as objectively over-valued, or to sell one that is perceived as objectively under-valued (e.g., if the agent expects the market imperfection that allowed the price to deviate from the objectively warranted value to persist for a sufficiently long period of time). What is methodologically important is whether such "fair" prices exist (or are perceived to exist) in the first place, and how they are formed (or perceived to be formed). In the next section of this chapter, I shall juxtapose the theoretical models that underlie EMH and RWH, and spell out the empirical assumptions involved. In the third section, I shall examine the veridicity of those assumptions as applied to actual capital markets. Finally, in the fourth section, I shall draw some methodological conclusions.

MARTINGALES

Since the 1960s, the prevalent way of explicating EMH has been by treating the expected returns on shares (i.e., expected dividends plus expected capital appreciation, normalised by the purchase price) as martingales. This means that the best bet on the rate of return in period t+1 is the rate of return that actually obtained in period t:

$$E(r_{t+1}|\Phi_t) = r_t,$$

where Φ_t is the information set available to the investor(s) in period t. This can be trivially reformulated thus: the difference between the rate of return in t+1 and in t is a *fair game* (i.e., with zero expected pay-off):

$$E(r_{t+1}-r_t|\Phi_t) = 0.^{134}$$

[134] The expectation of a difference equals the difference of expectations, hence $E(r_{t+1}-r_t|\Phi_t)=E(r_{t+1}|\Phi_t)-E(r_t|\Phi_t)$. Since Φ_t is the information available at time t, which on any definition includes r_t, $E(r_t|\Phi_t)=\text{const}=r_t$.

Another way of putting this is to say that future returns are unforecastable (since in the next period they are just as likely to be less as to be greater than current returns). It can be shown (although I do not do it here) that rates of return on shares are martingales **iff** share prices (inclusive of accumulated dividends[135]) are also martingales, or:

$$E(p_{t+1}|\Phi_t) = p_t,$$

and since according to EMH, $E(p_{t+1}|\Phi_t) = p^*_{t+1}$,

$$p^*_{t+1} = p_t,$$

and taking expectations on both sides and remembering again that $E(p_t|\Phi_t) = p^*_t$, we get

$$E(p^*_{t+1}|\Phi_t) = p^*_t.$$

The martingale model, however, does not simply explicate the original EMH. Samuelson (1965, 1973) proved that if betting on the change in the rate of return on a security (including shares, bonds etc.) is a fair game, then the market price of that security equals the present-discounted expected value of the cash streams to which it gives title, *as an identity*. In other words, the martingale/fair game model implies that at any given time, the price of a security equals its "real", or "fundamental", value:

$$p_t = p^*_t,$$

which is a lot stronger than the original assumption that $E(p_t) = p^*_t$. Thus, in the martingale model it is impossible to consistently outperform the market by buying underpriced shares and selling overpriced ones (neither over- nor underpriced shares exist to begin with). On the contrary, in the original form EMH presupposed that actual prices *fluctuate* around the fundamental value and thus it is possible to outperform the market by establishing, in the course of "fundamental analysis", which shares are "cheap" and which are "expensive".

[135] From now on, I shall use "p_t" to refer to price at *t* plus dividend accumulated between *t*-1 and *t*.

The martingale model is clearly a limiting case of the "fundamentalist" one. In the latter, the market price of a share is a random variable with a distribution whose mean equals the fundamental value of the share:

$$p_t = p^*_t + \epsilon_t, \text{ where } \epsilon_t \sim (0, \sigma^2).$$

If we further assume that the distribution has zero variance ($\sigma^2=0$), we are left with the martingale model (in which the empirically observed volatility of share prices exactly matches the underlying volatility of fundamental values). Hence, the possibility of fundamental analysis (and thus of outperforming the market) hinges on whether this particular restriction is justified.

But the zero-variance assumption is not the only one that can be imposed. It has been shown in recent literature (LeRoy 1989) that mathematically, the random walk model is merely a very restrictive version of the more general martingale model. In the martingale model, the fluctuation of share prices (=fundamental values) occurs with unrestricted variance (so that, e.g., positive or negative autocorrelation between share price variances at t and at $t+1$ is neither stipulated nor ruled out):

$$p_t = p_{t-1} + u_t,$$

where $E(u_t) = 0$. If we use "Ω" for the variance-covariance matrix of u_t, we can re-write the martingale model as

$$p_t \sim (p_{t-1}, \Omega).$$

In the random walk model, the share-price variance is supposed to be history-independent; that is

$$Cov(p_t, p_{t-1}) = 0.$$

This means that Ω is assumed to be diagonal (i.e., with all off-diagonal elements that correspond to cross-variances, equal to zero). Hence, it turns out that the original opposition between EMH and RWH can be formulated as that between an unrestricted and

a doubly restricted model. It means that we have a family of "nested" theories that are, in principle, capable of empirical test.

Before moving on to the empirical implications of the models considered above, we should introduce some complications. There are (e.g., in Fama 1970) three interpretations of what constitutes the relevant information Φ_t in the above formulas, and correspondingly, three versions of EMH:

(1) Only the information on past share price movements is included ("weak efficiency");

(2) All publicly accessible information is included ("semi-strong efficiency");

(3) All relevant, including privately held, information is included ("strong efficiency").

I shall argue that this classification is incomplete. It should be supplemented by a distinction between "actual efficiency" (the martingale model) and "expectational efficiency" (the fundamentalist model). (In fact, the distinction is clearly indicated e.g. in LeRoy; he speaks of the "strict version of market efficiency" referring to what I mean by "actual efficiency". I prefer different terminology to avoid potentially confusing and awkward expressions like "strict weak efficiency" and "lax strong efficiency"). I shall argue that market efficiency, especially strong efficiency, can be only expectational.

WHICH EFFICIENCY?

First of all, let us note that weak efficiency is not *prima facie* feasible. Indeed, it is well known that current and expected interest rates, anticipated inflation, unemployment and other macroeconomic data, as well as leaks on forthcoming mergers and take-over bids, changes in the regulatory regime and so on and so forth, heavily influence expected returns and share prices. Assume that the shares of two companies have had exactly the same price history since their launch. The only aspect in which the companies differ is that one supplies the domestic market and the other exports. An exogenous rise of the pound against foreign currencies would hurt the exporter but not the domestic supplier (the latter might, in fact, benefit from the decreasing costs of imported materials and from the downward pressure on the interest rates). Hence, the share price of the exporter will plummet and of the domestic supplier, most probably, increase. But this absolutely reasonable result would be inconsistent with weak efficiency, since the companies would remain identical with respect to the relevant information set (i.e., the history of past returns). The issue, therefore, is between semi-strong and strong efficiency, either of which can be cast in actual (martingale) or expectational (fundamentalist) terms.

The main assumption behind the martingale model, as pointed out in Samuelson (1965), is that individuals are risk-neutral. It means that they will be indifferent between two securities with the same fundamental values, even though the underlying probability distributions over the streams of future profits may be more level for one than for the other. In terms of market prices, only the mean of the price matters, not its variance. This rules out the existence of securities with negative expected returns, because the only possible justification for holding them would be as an insurance policy against a general downturn in the market. Let us suppose that a security is negatively correlated with the market, i.e., its expected growth is positive conditional on a market downturn and negative conditional on a market upturn. Then a risk-averse, but not a risk-neutral, agent will make such a security part of his portfolio in order to mitigate the volatility of the other holdings, even at the expense of somewhat diminished expected returns on the overall investment. Hence, with risk-aversion, such "insurance" stocks will be priced to yield negative expected returns, which is inconsistent with the martingale model.

The risk-neutrality assumption cannot be sustained. Indeed, economic theory often presupposes that whereas *individuals* are risk-averse (they prefer sure bets to gambles, and are prepared to pay an "insurance premium" by choosing a security with a lower expected value if that value has a sufficiently low variance), *institutions* are risk-neutral. And since institutional investors dominate capital markets, goes the argument, risk-neutrality is a valid supposition.

I strongly disagree with this. Firstly, the influence of individual investors would still be strong enough to distort the picture if the bulk of trading were conducted on a risk-neutral basis. But it is not; institutions are not risk-neutral either. This is confirmed, among other things, by the existence of the futures and options markets. These financial instruments have the function of smoothing out the fluctuations in securities prices and exchange rates. Many insurance and investment companies are themselves insured against excessive losses (e.g., in the Lloyds of London it is individual underwriters who bear the bulk of the risk in exchange for exceptionally high expected returns).

A fairly recent phenomenon of "funds of funds" where an investment trust invests not in the shares of conventional companies but in other investment trusts, thus further spreading the risks, is another example. Finally, pre-election and other uncertainty is often blamed for sluggish performance of markets. There is a simple explanation why institutional investors should be risk-averse: they are ultimately owned by individuals, and although the pooling of resources allows institutions to take a longer-term view of investing (and thus to take a less risk-averse point of view compared to individuals), the difference is one of degree only. So, the martingale model, which relies crucially on the assumption of risk-neutrality, must be empirically wrong.

However, this conclusion is of limited utility. It is common knowledge that *any* model is empirically wrong in one way or another and therefore distorts reality. The question is whether its assumptions are *so* wrong that the model distorts reality to an *unacceptable* degree. This is what I turn to next, by looking at the observed behaviour of prices. That behaviour seems to be inconsistent either with the unrestricted martingale model, or with random walk. Random walk is especially vulnerable to recalcitrant evidence. As is obvious, random walk implies that the sign of Δp_{t+1} is independent of that of Δp_t. However,

substantial empirical evidence contradicts that: two consecutive price changes of the same sign (continuations) are significantly less probable than two reversals, as documented in Niederhoffer and Osborne (1966). The unrestricted version of martingales has its own empirical difficulties, as one would expect given the implausibility of its underlying assumption of risk-neutrality. A huge body of research indicates that the volatility of actual share prices is greater than that predicted by martingales. Let us look at the argument more closely.

LeRoy and Porter (1981, also LeRoy 1989) prove a so-called "variance-bounds" theorem. They begin by splitting actual returns into the rationally expected and the unexpected components. The unexpected component of a one-period return can be defined as

$$e_t = p_t + d_t - E_{t-1}(p_t + d_t),$$

with all the right hand-side variables same as before (it is assumed that e_t's mean and variance are finite). If all the future unexpected elements of actual returns are summed up over infinity

$$x_t = \Sigma_{i=1}(1+\rho)^{-i} e_{t+i},$$

it would be easy to show that

$$p^*_t = p_t + x_t,$$

i.e., at any time, the fair (in LeRoy's expression, "ex post rational") price equals its actual value plus the sum of all future unexplained deviations. And since actual prices are what the market predicts the fair prices will turn out to be, based on the information available, the fair price equals the market forecast plus the forecast error. Assuming that market forecasts are "optimal" (i.e., contain no systematic biases), forecasts themselves and forecast errors are contemporaneously independent:

$$Cov(p_t, x_t) = 0.$$

This uncorrelatedness implies that

$$V(p^*_t) = V(p_t)+V(x_t),$$

and since variances are by definition non-negative, $V(p^*_t)$ is an upper bound for $V(p_t)$, hence the name of the theorem. So, the martingale model implies that $V(p^*_t) = V(p_t)$, which makes it empirically testable. Unfortunately, this theoretical prediction is reversed in real-life asset markets. *Pace* the martingale model, actual returns are more volatile than expected ones. E.g., Campbell and Shiller (1988) report that actual prices have twice the standard deviation of those implied by the martingale model.[136] This indicates that the model is flawed but does not yet tell us what is wrong. It could be the underlying EMH, or else the additional restrictions imposed on it by martingales. To answer this question, some additional analysis is in order. Among LeRoy and Porter's results was that $V(p^*_t|\Phi_t)$ is independent of Φ_t. This, in turn, implies that $V(p_t)$ and $V(r_t)$ are inversely related:

> That is, the more information agents have, the higher is the variance of price and the lower is the variance of returns. Thus if agents have very little information, stock prices are usually not much different from the discounted sum of unconditional expected dividends, a constant. Therefore stock prices have low volatility. In this case realizations of actual dividends come as near-complete surprises, inducing high volatility in actual returns. However, if agents have a great deal of information about future dividends, stock prices have almost as much volatility as discounted actual dividends, the two being highly correlated. In this case significant surprises occur very seldom, implying that returns will usually be nearly equal to their unconditional expectation (LeRoy 1989, 1697-8).

What can we make of it? First of all, this indicates that the agents operating in capital markets utilise a lot more information than just the history of returns. If that were all, stock prices would have low volatility, which is not the case. In order even for the equality of variances $V(p^*_t)=V(p_t)$ to hold, the stock of data Φ would have to be fairly rich, i.e., to include at least all the relevant publicly-available information.. But for the variance of actual prices to actually exceed that of rationally expected ones, $V(p_t)>V(p^*_t)$, the

---

[136] It must be clear that share price volatility has entered the discussion from two very different angles:

(a) a negative correlation between *actual* price changes in any two *subsequent* periods, which is inconsistent with RWH; and

(b) an *excess* variance of *actual* compared to ex ante *rational* prices (over *the same* period), which is inconsistent with martingales.

assumption of zero-correlation between predicted returns and prediction errors must be dropped in favour of negative correlation: $Cov(p_t,x_t)<0$. This implies that future returns are partially forecastable and *not* a fair game as postulated by the martingale model. In other words, we can safely assume that high estimates of future returns based on the current return will "overshoot", and low estimates - "undershoot" the mark. This conclusion fits in well with the widely accepted observation that capital markets overreact to price movements. A possible explanation of this phenomenon is that market agents, correctly anticipating the direction and approximate magnitude of a price change, amplify that change in the stampede to benefit from the development. In extreme cases, such behaviour leads to so-called "speculative bubbles" where the price of an asset is expected to rise fast, and the expectation becomes a self-fulfilling prophecy until the price sky-rockets to an absolutely unsustainable level whereupon the bubble bursts and the price plummets far below the original (and approximately "correct") level. An interesting corollary of this analysis is that speculative bubbles are qualitatively identical with more "normal" patterns of market behaviour.

I hasten to point out that a negative correlation described above, while inconsistent with martingales, leaves the core of the EMH (i.e., the thesis that actual prices are based on "fair" fundamental values which, in turn, depend on expected future profits) quite intact. Even though the distribution of actual prices around the fundamental value is leptokurtic relative to the normal (i.e., thicker in the tails and thinner in the middle), nothing suggests that its mean is not the fundamental value itself.

To sum up, evidence suggests that the restricted (martingale) and, *a fortiori,* the doubly restricted (random walk) versions of market efficiency are untenable, which leaves the fundamentalist model. However, we are entitled to be dissatisfied with this conclusion. Originally, market efficiency (in its fundamentalist form) and random walk were conceived of as two contrary models of capital markets, in the same way as the wave theory of Huyghens and the corpuscular theory of Newton were contrary models of optical phenomena. Now we are dealing with one theory (RW) superimplying (or being a proper subset of) another (EMH). What if the whole EMH framework is wrong? Does a non-martingale (fundamentalist) version of EMH have independent support?

TIME IS OF THE ESSENCE

One may have noticed by now that there is a discrepancy between two kinds of empirical evidence relating to market efficiency. One is the evidence indicating that returns are non-forecastable (which was one of the major original objections against the fundamentalist model). The other is the evidence of excess price volatility which strongly suggests that returns *are* partially forecastable. Surely returns cannot be both? Not unless we are measuring them over different time horizons. As LeRoy (1989) points out, the first sort of evidence is based on short-term observations (one-day and one-week returns). The second sort, however, is based on price movements over the medium term (more than a year). LeRoy refers to Fama and French (1988) who looked at the correlation between the average returns on a security in periods between t-T and t, and between t and t+T, and regressed that correlation on the length of T. They found zero correlation for T up to one year; the correlation coefficient then peaked at 0.35 for T between 3 and 5 years, and fell back to zero for T of ten years. In other words, returns are partially forecastable in the medium term but non-forecastable in the short- and long term.

I have mentioned above that, by the variance-bounds theorem, complete unforecastability of returns is equivalent to an exact identity between the actual price and the fundamental value. Either expresses the economic gist of the martingale model, and both crucially depend on the assumption of risk-neutrality. Hence, the discovery that returns are partially forecastable, especially in the medium term, gives strong support to the "fundamentalist" ("expectational") version of market efficiency. "Partially" is important here, for if the degree of forecastability had been much higher, efficiency would have been out of the question.[137]

Assuming that all individuals hold their assets willingly and are risk-averse, it must be possible, in theory, to calculate their probability distributions over future returns that reflect the risk premium they perceive to be present in their portfolios. Such probabilities will

[137] *Some* forecastability is OK; it would seem to lead to efficiency-defying arbitrage possibilities (i.e., supra-natural profits at no extra risk), but the latter are hugely diminished by the extra costs involved in fundamental analysis-based active investment, compared to the passive buy-and-hold strategy: brokerage and other transaction costs, opportunity costs of taking the time to manage the portfolio, etc.

vary from individual to individual, since they have different degrees of relative risk aversion but nevertheless find it marginally attractive to hold the same stock (although for every buyer there is a seller, every single stock is simultaneously held by numerous market agents; moreover, the seller finds a stock only marginally less attractive than the buyer, otherwise the former would have sold, and the latter would have bought, the stock even earlier).

This seems to me to favour strong efficiency, since an exclusion of private information from the information set Φ_t would lead, provided Φ_t is sufficiently large for convergence, to virtually identical probability distributions, and, assuming varying relative risk aversion among individuals, to market failure: some stocks would find no buyers and others - no sellers. The presence of private information in Φ_t guarantees that convergence of probabilities does not occur but convergence in subjective valuations (i.e., with the risk element accounted for) does.[138] If, on the contrary, we were to assume risk-neutrality, the market would clear only if all subjective probabilities were quasi-identical.[139] Hence, the martingale model, were it empirically acceptable, could only be cast in the semi-strong form.

An interesting insight into the issue is afforded by Robert Lucas's (1978) approach which utilises the idea that individuals maximise the present-discounted sum of future utilities (which themselves are a function of consumption in each period). In his model, the marginal utility of consumption may vary over time, and equilibrium prices are such that for every individual,

$$p_t U'_t = (1+\rho)^{-1} E_t(p_{t+1}+d_{t+1})U'_{t+1},$$

which reduces to the martingale model if we assume risk-neutrality. Indeed, under risk-neutrality the utility function is linear, therefore marginal utility is constant ($U'_t=U'_{t+1}$ for all

[138] Donald Gillies suggests that restriction to publicly accessible information would do the trick, since there is so much of it that no two individuals would actively use in their calculations an exactly identical sub-set of public data. This illustrates the fuzzy boundary between what is publicly accessible but, in fact, accessed only by few, and what is private.

[139] Such hypothetical "universal" probabilities are appropriately called "risk-neutral" in finance literature.

t), and both sides can be divided by U'. This again shows that martingales are intrinsically related to risk-neutrality but are not restricted to it: if relative risk-aversion is positive but constant over time, the U' terms again cancel out and we still get martingales. Constant risk aversion is a restrictive assumption, but not nearly as restrictive as risk-neutrality.

To sum up, although the current state of empirical and theoretical research leaves a number of loose ends, "expectational" market efficiency comes across as the best explanation of the behaviour of capital markets, and in the foundation of EMH is the idea that agents use subjective probability and utility functions to create rational expectations of future returns, which they update according to the rule of conditional probability as their information set evolves. This conclusion undoubtedly strengthens the general claims of subjective Bayesianism; and in my view, this support far outweighs the odd violation of transitivity cooked up in a psychology lab.

APPENDICES

CARTWRIGHT ON "STATISTICAL DISCOVERY"

Introduction

Nancy Cartwright's view that capacities (or, in more widely accepted terminology, "propensities" or "dispositions") are essential (perhaps even principal) building blocks for scientific ontology is well known (cf. Cartwright 1989). This view includes the idea that statistical, observation-level, regularities, are best thought of as based on the underlying foundation of capacities, and that properly conducted, observation-level generalisations can result in discovering such fundamental dispositional structures. Cartwright's view is not original; e.g., Rom Harré has been a vocal advocate of the "dispositional" philosophy of science for at least thirty years. However, Cartwright's version of dispositionalism has been in the focus of philosophical discussion owing to her forceful attempt to relate the general methodological issues involved to their applications in special sciences, including economics. I happen to agree that dispositions are extremely important in science, both for intellectual coherence and methodologically. It does not mean, however, that Cartwright's analysis of the relation between dispositions and their statistical manifestations in economics does not require scrutiny.

The need for extra caution becomes apparent when one realises that her acquaintance with economics and econometrics is mostly second-hand. In Cartwright 1989, virtually all references to econometric literature are taken from Mary Morgan's (1989) *The History of Econometric Ideas*, and the superficiality of Cartwright's knowledge of economic theory proper is quite striking. For instance, as an example of a tension between economic theory and econometric practice she considers the (linear) relationship between demand q and price p:

$$q = \alpha p + u,$$

where u is the error term:

> Consider α then: α represents the price elasticity of demand, a concept that was significant in the marginal revolution. But to my mind, α is treated very differently by the econometricians from the

> way it was treated, for example, by Alfred Marshall. When Marshall introduced the idea of demand elasticity in his Principles of Economics, he immediately proceeded to consider 'the general law of the variation of the elasticity of demand'. How does the elasticity change? For instance, the price elasticity of demand will itself depend on the level of the price; the elasticity of demand is great for high prices, and great, or at least considerable, for medium prices; but it declines as the price falls; and gradually fades away 'if the fall goes so far that the satiety level is reached'. In contrast, the econometricians by contrast treated the elasticity as if it did not vary: α measures an abiding or stable tendency of the system [...] What I want to stress is that the assumption of stability is built into the demand equation. From the discussion of Marshall one would expect that α was a function of p, i.e. $\alpha = \alpha(p)$, and of a number of other features as well: $\alpha = \alpha(p,...)$. But that is not how it is treated. It is treated as a constant - or fixed - parameter (Cartwright 1989, 150-1).

How real is this contradiction? Contrary to Cartwright, α in the above equation is *not* the price elasticity of demand; $\alpha p/q$ is. Therefore, there is not a shred of inconsistency that Cartwright perceives here: α is indeed stable, just as the econometricians think (or rather, hope), while the demand elasticity varies depending on the p/q ratio, that is, on the point at which the measurement is taken.[140] Not to mention that both microeconomics and econometrics require, for their independent reasons, that the demand function $q = q(p)$ have an intercept term which would represent the level of demand at zero price. As the demand function is represented in Cartwright, no proper regression could be run. Having made the source of some of my reservations clear, I proceed to look at her substantive claims.

Cartwright's main contention is that formal econometric methods constitute the mechanism whereby economists discover the underlying dispositional structure of their subject area, given sufficient theoretical assumptions:

> ...given the kinds of very strong assumptions that go into causal models, it is possible to extract causal information from statistics (Cartwright 1989, 13).

So far, rather innocuous if weak. But later she unpacks this statement, revealing its composite nature:

---

[140] Some economists and especially econometricians do, occasionally, loosely refer to α as "elasticity", which is what must have confused Cartwright. When pushed, though, they will readily revert to the universally accepted definition: price elasticity of demand $e_{q,p} \equiv \partial q/\partial p \cdot p/q$.

> The arguments of this chapter were meant to establish that the inference from probability to cause can have the requisite bootstrap structure: given the background information, the desired causal conclusions will follow deductively from the probabilities. In the language of the Introduction, probabilities can measure causes; they can serve as a perfectly reliable instrument (Cartwright 1989, 35).

Concerning the first thesis ("...the desired causal conclusions will follow deductively from the probabilities"), I shall argue that econometrics is capable of nothing of this sort. On the contrary, as is clear from its name, it merely allows a measurement of the adjustable parameters of independently postulated economic models. In other words, statistics in general and econometrics in particular merely accept that the causal relations postulated by a theory exist, and measures their strength. As to the second thesis ("...probabilities can measure causes" in a "perfectly reliable" way), it becomes quite acceptable once the "perfectly reliable" bit is removed. It appears, however, that for Cartwright these two claims are if not identical, then inextricably related. Why it should be so, is not immediately clear. The first thesis is much stronger; surely measurement and discovery are only tenuously related. It is one thing to measure something that we know is there, and another - to find out whether or not it *is* there. Therefore, I shall concentrate on the "discovery claim" before addressing the "perfect reliability" version of the "measurement claim".

The "statistical discovery" thesis

What kind of 'very strong assumptions' would, in Cartwright's view, help reveal the causal structure of a system under study? Using her catch phrase "No causes in, no causes out", one suspects that in order to *deduce* a causal structure, one has to *postulate* a causal structure first. Clearly, the postulated structure cannot be the same one that we seek to discover, otherwise what would be the point of statistical inference? So, it must be *some other* causal structure (let us call it the "input structure") from which the "output structure" must be deductively derivable by statistical inference. *A priori*, this "boot-strapping" (Clark Glymour's term adopted by Cartwright) approach looks extremely implausible. If something is deductively derivable from a set of assumptions, then no extra inferential mechanism, on top of deductive logic (possibly with an expanded set of axioms and/or rules of inference), is required. If, on the other hand, the conclusions do not follow from

the premises by predicate calculus (say), it is hard to see how the addition of the axioms of probability and other tools of statistical reasoning would help. After all, statistics was invented specifically to handle those situations where no certain conclusions can be derived. If one is a Bayesian, its conclusions are of the form "On evidence *e*, hypothesis *h* is true with probability *p*", and unless *e* actually entails *h*, in which case no statistical techniques are needed, *p* will always be less than one. And if you are not a Bayesian, still no determinate causal conclusions are forthcoming. My guess is that Cartwright became seduced by the apparent plausibility of Glymour's approach as applied to deterministic (non-statistical) episodes from the history of physics, such as Newton's famous "deduction" of the composite nature of light (where, indeed, the "crucial experiment" of the *Principia*, in conjunction with the background assumptions, rules out the rival theories). The decisive difference between deterministic and statistical models, in this respect, is that in the latter, the set of theoretical alternatives is usually infinite or even non-denumerable, hence deduction by elimination cannot possibly succeed

Let us look at Cartwright's argument more closely. She utilises J. L. Mackie's notion of "*inus* conditions" ("*i*nsufficient but *n*on-redundant part of an *u*nnecessary but *s*ufficient condition" – Mackie 1980, 62). An example of an inus condition for the production of a fire would be lighting a match. It is insufficient because the presence of oxygen and of a flammable substance are also required, but it is non-redundant since without it a fire would not start either. The whole "burning match-oxygen-flammable substance" set-up is clearly sufficient but unnecessary (oxygen could be substituted with a halogen gas, or the match - with an electric sparkle or nothing at all; lithium will burn at room temperature when exposed to air). How does the notion of an inus condition help us in discussing the issue at hand? If we use X_i to designate an inus condition and A_i - the completing background conditions which, jointly with X_i, constitute an unnecessary but sufficient condition A_iX_i, then the full causal structure E can be described as the disjunction of all such conjunctions:

$$E \equiv_{def} \mathrm{V}_{i=1} A_i X_i.$$

Mill calls the right-hand side of the identity the "full cause", and Cartwright labels any conjunction of X_is a "full set". Such "full sets" enter as regressors into the right-hand side of regression equations. Mackie made a point of emphasising that an inus condition may

well be a spurious cause if the correlation between the condition and its putative effect is either purely accidental or due to a common cause. For instance, suppose that the sounding of a hooter at 5pm in Manchester is highly correlated with the stoppage of work in London. Surely we would not want to treat one as the cause of the other. Rather, a common social convention at work in both cities might produce the correlation. The task of econometrics is to help us separate such correlations from genuine causal connections.

Now Cartwright's dictum "No causes in, no causes out" appears to mean that an econometrician postulates, based on economic theory, that a number of inus conditions must be tentatively included in a regression equation, and having run the regression, eliminates those which prove to be spurious. Thus "deduction by elimination" is supposed to work in econometrics. The background conditions mentioned in the above description of Glymour's bootstrapping method, provide the set of originally included regressors. Those of them that subsequently show no "genuine" correlation with the dependent variable are chucked out, and thus the true causal structure is revealed.

Unfortunately, this picture is exceedingly naive. First of all, an absence of a causal correlation might be revealed in different ways.

(1) The coefficient obtained in a regression may be very small (close to zero). But this, by itself, does not allow us to conclude anything, as the rule telling us to disregard putative causal factors whose regression coefficients are close to zero, is both too strict and too lax. Surely some genuine causal factors are more strongly related to the effect than others, hence the absolute values of regression coefficients will vary widely. There is no natural "demarcation line" between the regression coefficients of genuine and spurious causes. Besides, a simple re-scaling will turn a low coefficient into a high one, or vice versa, as long as it is not *exactly* zero to begin with. But measure-theoretically, such a possibility can be safely disregarded as it has probability zero.

Alternatively, we could try to use as the criterion of demarcation the "Pearsonian correlation coefficient"

$$r_i = \Sigma x_i y / n s_{xi} s_y,$$

where s_{xi} and s_y are the sample means of the *i*th regressor and of the dependent variable, x_i and y - the deviations of the regressor and of the dependent variable from s_{xi} and s_y respectively, and n - the sample size. Unlike the regression coefficient, r is scale-invariant. Still, we are nowhere closer to a non-arbitrary cut-off point that would separate spurious correlation from (relatively weak) genuinely causal ones. Also, r measures only *linear* correlation, so even if a certain factor is a very strong causal determinant but its relation to the effect is non-linear (as, e.g., between output and average variable cost in firms' production decisions), r will be close to zero. At the same time, some very high correlations have been found between clearly causally unrelated variables, such as money supply in England and the rainfall in Scotland. Admittedly, Cartwright could argue that such outlandish connections would not be part of the "input causal structure" and hence would not count as counter-examples. It is not clear, however, why such high degrees of correlation might not arise between factors that *are* tentatively admitted on the "input side" but turn out to be causally unrelated, after all.

Finally, we could judge a putative causal factor by its contribution to the degree of fit of the regressed equation to the data from which the regression coefficients were derived. First of all, it must be noted that since *any* extra regressor, no matter how unlikely, improves the goodness of fit, we cannot simply retain all those regressors that lead to such an improvement. We could decide to chuck those regressors which do not *significantly* increase the goodness of fit. Again, there is no natural threshold separating "significant" from "insignificant" increases. An increase in the goodness of fit may be small if the extra regressor is only slightly, but nevertheless genuinely, causally related to the quantity which is being regressed. This would be expected, for instance, if we included the exchange rate of the pound sterling to the Spanish peseta into a regression for the demand for sun-glasses in the UK.

Otherwise, we could try to base our assessment of the causal import of a factor not on sample statistics, but on our best estimates of the population statistics. This, for instance, would allow us to conjecture that the regression coefficient, for the infinite population, is zero, thus providing a clear-cut demarcation criterion. Unfortunately, since the population statistics are, in general, not accessible, we have to settle for (fallible) estimates. Suppose that we conjecture that a certain factor has no causal relation to the effect under study, and

therefore the "true" population regression coefficient can be expected to be zero. There are a variety of different tests available. For instance, in the Likelihood Ratio Test we test the hypothesis $\alpha_i=0$ by running a regression on the restricted model from which the regressor x_i is excluded, and the unrestricted model. The ratio of their likelihoods (to calculate which we need to know the distribution of the error term) will show how much violence to the truth is done by the restriction $\alpha_i=0$. The ratio ranges from 0 to 1; if the ratio is only slightly less than 1, then we may conclude that the omitted regressor is relatively unimportant. More formally, based on the assumed distribution of the error term, the relevant test statistic will be a particular function of the likelihood ratio and distributed as chi-squared with one degree of freedom (assuming that we are testing only one restriction at a time), yielding a particular probability of the restriction being true.

Alternatively, one could use the Wald Test (where the test statistic is based on the unrestricted model only) or the Lagrange Multiplier Test (where the test statistic is based on the restricted model only). In each case, although the test statistic is different, it is chi-squared distributed with the degrees of freedom equal to the number of restrictions, and the probabilities derived in the alternative tests converge to the common value as the sample size increases. What is important, we never get anything other than a *probability*, not a certainty, that a certain factor is causally salient.

(2) The coefficient obtained in a regression may be fairly high. It does not yet guarantee that the regressor corresponds to a genuine cause, as Mackie's example of 5pm work stoppage illustrates. Cartwright looks at Reichenbach's Principle of the Common Cause that can be rephrased thus: every inus condition that displays a high degree of correlation with its putative effect but does not embody a genuine causal relation with it, must have a common cause with the effect. Cartwright takes a more general approach:

> There are a variety of other causal stories that could equally well account for the correlation. For example, the cause of one factor may be associated with the absence of a preventative of the other, a correlation which itself may have some more complicated account than the operation of a simple joint cause. More plausibly Reichenbach's principle should read: every correlation has a causal explanation. [...] The simplest view would be this: if a factor is an inus condition and yet it is not a genuine cause, there must be some further causal story that accounts for why it is an inus condition.

> The argument above uses just this idea in a more concrete form: any formula which gives a full set of inus conditions, not all of which are genuine causes, must be derivable from formulae which do represent only genuine causes (Cartwright 1989, 29).

But this condition is simultaneously too weak and too strong. It is too weak because, as we have seen, formulae which give full sets of inus conditions are disjunctions, and as such, are trivially derivable from shorter disjunctions that include only those disjuncts that represent genuine causes. It is too strong because, without any justification, it rules out purely coincidental correlations. It is not difficult to see a common causal background in Mackie's Manchester/London example. But one would not even know where to start looking for a causal story relating the vagaries of Scottish weather and the monetary policies of the Bank of England.

Also, the mechanism of "deduction by elimination" suggested by Cartwright's picture, does not square with stories more complex than the "common cause" scenario considered above under (1), because such more complex stories cannot possibly transpire by eliminating one of the constituents of a disjunction. In fact, Cartwright's drawing our attention to such more complex stories completely undermines her own "statistical discovery" thesis.

Let us assume now that an econometric model contains a spurious regressor. How would we eliminate it? If we just drop it, the degree of fit between the model and the data will decrease, possibly substantially. If the decrease is small, we still have all the "threshold" problems described in (1). Also, the only reason why it would be small is that the removed regressor is highly correlated with one of the remaining ones. In other words, our model contains *multicollinearity*. In such a case, we would have faced identification problems in the first place (the variance of the regression coefficients would have been obscenely high and the whole regression would have been useless for prediction and even systematisation). Additionally, *which* of linearly dependent regressors to remove, is a matter of theory-educated judgement and cannot be decided on formally-statistical grounds.

If the decrease in the degree of fit associated with the removal of a spurious regressor is large, we have other problems. Clearly, compromising the empirical adequacy of a model

just because we *a priori* do not believe that some part of it is causally relevant, is not very satisfactory. We should decide what causally relevant factor the spurious regressor stands for as a proxy. If it was not in the model to begin with, its discovery will have nothing to do with the statistical testing procedure as such; it will be a theoretical decision. Moreover, the causally relevant factor might not be quantifiable unlike its causally impotent proxy. Imagine that we want to explain an artist's posthumous popularity as measured by the number of reproductions of his paintings printed annually, the number of people attending his exhibitions, etc.. Clearly, the artist's talent would be a very important causal factor, but it is not measurable. We would have to use a proxy variable such as, e.g., the number of his paintings sold within his lifetime and up to twenty (say) years after his death (to correct for the buying public's conservatism). In such a case removing the proxy would not only be statistically unjustified, but also empirically undesirable, although keeping it would obscure the true causal picture.

In other words, no formal measure would allow the econometrician to eliminate definitively all causal structures but one. Cartwright's view presupposes that we have an exhaustive statistical model of the subject area that, usually, contains redundant and/or wrong elements. The latter are then "weeded out" using formal statistical techniques. But in fact, we might as well not have the causally relevant regressors to begin with, and no amount of statistical inference will tell us what they are (there are sometimes certain pointers, to be sure: e.g., heteroskedasticity, or non-constant variance of the error term, usually indicates autocorrelation, i.e., the dependence of the endogenous variable on its own past values; however, in most cases such simple rules of thumb do not exist). Hence, Cartwright's trust that econometrics would allow, given enough background assumptions, to deduce the true causal or dispositional picture of the subject area, is completely misplaced.

Is a weaker thesis defensible? If we cannot "discover" a causal structure by purely statistical methods, can we at least *prove* (or disprove) that a given, well-established relationship, is a causal one? Let us consider an example of how Cartwright sees such a procedure. She takes from Kevin Hoover's article (Hoover 1990) a two-variable model in which "money causes price". Remembering that money does not *cause* price (implicit prices of goods exist in moneyless economies), we should rephrase this awkward and

misleading expression, e.g., "the nominal money supply M influences the level of prices P". The model is:

$$M = v$$

$$P = \alpha M + \xi,$$

where v and ξ are random error terms. What is the best estimate of α? Cartwright makes a overly ambitious statement that

> In the nice case when v and ξ are uncorrelated, α can be identified: $\alpha = \langle MP \rangle / \langle M^2 \rangle$ (Cartwright 1995, 66).

$\langle MP \rangle$ and $\langle M^2 \rangle$ are the sample covariance of M and P and the sample variance of M, respectively. In fact, α can only be (fallibly) *estimated*, not *identified*, and the Ordinary Least Squares (OLS) estimator $a = \langle MP \rangle / \langle M^2 \rangle$ is only one possibility, admittedly the most commonly chosen one. This may look like nit-picking, but Cartwright's talk of "identifying" unknown population parameters illustrates the naive optimism with which she views econometric inference. Getting to the substance,

> The surest way to find out about the causal connection between M and P is to conduct a controlled experiment. The experiment requires a control variable that can be used to change the distribution in M. If the distribution in P changes correspondingly, we conclude that M causes P; if not, in the simple two-variable case we conclude that M does not cause P. Two characteristics of the control variable are crucial. It must satisfy two important conditions: the control variable must be a known cause of M and we must be assured that it varies independently from all causes of P operating at the same time except M (Cartwright 1995, 66).

In order for the second condition to hold, says Cartwright, v and ξ "must stand for causes of M and P, respectively". This is completely untrue. To begin with, the first condition (that the control variable be a known cause of M) is too strong; they may be identically equivalent without one causing the other, as e.g. total income in a closed economy without a government sector is *defined* as the sum of consumption and investment:

$$Y \equiv C + I.$$

In such a case the left-hand side will perfectly co-vary with the right-hand side while the relationship between them is purely conceptual, not causal. (In general, far from all *bona fide* correlations are causal.)

The second condition, that v and ξ be independent, is desirable but not essential. It is desirable because if it holds, the Maximum-Likelihood (ML) estimator of α will coincide with the OLS estimator, provided that the distribution of ξ is known to be normal, and the estimation process will be much more straightforward than otherwise. But it is not essential because α could still be estimated, although by more complex methods, even if v and ξ are not independent. Further, even for their independence it is not necessary, *pace* Cartwright, that they represent causes of M and P, respectively.

Moreover, it would be an *exception* if, given a presumed causal relationship between M and P, their own causes were uncorrelated. Of course, we could guarantee that independence *by construction*, as Cartwright acknowledges, by bundling into ξ all those causal (and, I should add, relevant non-causal) determinants of P that are *uncorrelated* with M (and therefore, by stipulation, with v). The point that Cartwright fails to appreciate is that the independence of v and ξ could be much more easily guaranteed if they were *not* the causes of M and P, respectively

For instance, regressing the annual number of bankruptcies in the US economy on the interest rate on Federal Reserve bonds (a highly plausible causal relationship), we would not expect any correlation between the annual number of National Hockey League wins by the New York Rangers and the average yields of wheat in Minnesota - exactly *because* the Rangers' performance is not a plausible cause of the annual number of bankruptcies, and Minnesota wheat yields are not a plausible cause of the interest rate levels.

One might wonder why the causally impotent factors have been included in our equations in the first place, but that is not at issue. If A is regressed on B and C, while B itself - on D, it may turn out, *post*-regression, that A is independent of C and A - of D, and that *explains* why C is independent of D. However, that latter independence, albeit initially unexplained, may have been *presupposed* in order to run an OLS regression.

Anyway, what is the conclusion that Cartwright draws from the set-up described above? She argues as follows: M is a cause of P **iff** the distribution of the latter will change if we change the distribution of the former while keeping all other parameters constant. A change in the distribution of M will be effected by manipulating with v (the "control variable"), and since v and ξ are presumed independent, all the change in the distribution of P can be safely ascribed to the putative causal relationship between M and P. This would constitute a "controlled experiment" analogous to one in natural sciences.

However, continues Cartwright (possibly motivated by an awareness of the extreme difficulty of conducting controlled experiments in social sciences), such an experiment would be redundant anyway. Indeed, since we "know" that $\alpha=\langle MP\rangle/\langle M^2\rangle \neq 0$ **iff** $\langle MP\rangle \neq 0$, we also "know" *a priori* that a change in the distribution of M would affect that of P. Hence, a value of $\alpha \neq 0$ found from the sample statistics, guarantees that M is a cause of P.

But the value of α calculated as above is just an *estimate*, I would even say an *educated guess*, and therefore may well change as new measurements become available. Even if $\langle MP\rangle \neq 0$ in a finite sample, it may well be zero in the population, or vice versa. The point of gathering new data, be they generated by a natural process or through a controlled experiment, is exactly to assess the reliability of our estimates of regression coefficients and/or to correct them.

But Cartwright's conclusion that controlled experimentation is redundant, is faulty in more than one way. Suppose that we have observed that a change in the distribution of M affects the distribution of P, but not the other way around. Under certain assumptions, we can make a qualified conclusion that if there is a genuine causal relationship between the two, its direction is from M to P and not *vice versa*. Simply observing their covariance in the given data, on the contrary, tells us nothing about the direction of causality. Hence, Cartwright's view that statistical inference as such, given the right kinds of assumptions, is capable of discovering causal relationships, is fundamentally wrong even in its weak interpretation. Out goes the "discovery of causal structures" thesis. I now turn to the "perfect reliability of measurement" claim.

The "perfect reliability of measurement" thesis

I begin by repeating that it seems to me uncontroversial that in a certain sense, probabilities can measure causes (or rather, dispositions). What is doubtful is whether such a measurement can ever be very accurate, let alone "perfectly reliable".

> If correlations can be dictated by the laws of nature independently of the causal processes that obtain, probabilities will give no clue to causes (Cartwright 1989, 29).

This pessimistic conclusion, however, is a pure *non sequitur*. Indeed, it is much more probable that a particular correlation should occur if there *is* a causal story behind it than otherwise, but it is a matter of degree only. The existence of "purely coincidental" correlations merely illustrates that very unlikely states of affairs do obtain, and therefore the clues that probabilities give to causes are only *somewhat*-reliable *clues*, not iron-clad *guarantees*. It is worth reiterating that if we measure the strength of a causal factor by the coefficient attached to the corresponding regressor, "perfect reliability" goes out of the window too, since each finite-sample estimate of a regression coefficient is infinitely more likely than not to differ from the "true" (or "population") value.

THE COLLAPSE OF LTCM

Introduction

Long Term Capital Management, or LTCM, was something of a misnomer. Despite its name, it attempted to predict securities price movements in the short term. Therefore, its 1998 collapse can be viewed as evidence against the predictability of returns thesis. Although the Efficient Market Hypothesis chapter deals with the pricing of shares, and LTCM was primarily involved in fixed-income securities such as bonds, there is enough similarity between the two classes of securities to draw methodological implications from the LTCM affair. After all, the underlying issue at stake is whether, and how quickly, market forces 'iron out' the imperfections which make above-average returns possible, notwithstanding the class of assets used.

Hedge funds

LTCM, founded in 1994, was one of a growing number of *hedge funds*. A hedge fund is a private investment partnership that invests in a variety of securities. The use of the word 'hedge' implies the practice of 'hedging', i.e., managing risk through derivatives (such as futures and options). However, not all hedge funds invest in derivatives. Each fund is free to invest in any type of securities, depending on its trading strategy. Another common misconception about hedge funds is that their operations always involve high leverage (the ratio of total invested funds to own funds). To put simply, high leverage means high borrowing. While certainly true of LTCM, high leverage is not a necessary characteristic of a hedge fund: some of them do not use leverage at all, which others do not exceed the 2:1 ratio (i.e., invest borrowed and own funds in the same proportion). Thus, the genuinely distinctive feature of hedge funds (differentiating them from other pooled investment vehicles such as mutual funds) is their private nature. They tend to be owned by a small number of wealthy individuals seeking above-average returns and therefore willing to take above-average risks. Related to that is their secretive nature: they are not allowed to advertise, but are not required to disclose their accounts, either.

Having said that, a significant minority of hedge funds (including LTCM) do employ high leverage. For those that do, this is not a matter of choice but a result of their

investment strategy. The high-leverage funds tend to be the 'arbitrage funds' whose goal is to exploit transient market inefficiencies.[141] In general, such an inefficiency will take the form of fundamentally identical assets being priced differently. The arbitrageur will buy the cheaper of the two assets, hoping to sell it when it eventually appreciates to its 'true' value. The potential profit can be doubled at a stroke if the arbitrageur finances the purchase by 'short-selling' the more expensive one. Short-selling is the term for selling what you do not yet have. It is a feasible strategy if, as is often the case in the financial markets, delivery is at a later date than the transaction. (The arbitrageur hopes that, when it comes to delivery, the good can be bought more cheaply than at present). In due time, the reasoning goes, other market participants will realise that there is no good reason for the price differential, and the two prices will converge. Then the arbitrageur will have made a profit by selling, more expensively, the asset that used to be cheaper, and buying, more cheaply, the asset that used to be more expensive.

The 'Asian flu'

A big part of LTCM's investment portfolio was derivatives (futures and options) linked to several emerging markets in Asia, Eastern Europe and Latin America. The fund's strategy was based on the assumption that those markets would continue rising. However, reality went contrary to expectation. In late 1997, several Asian markets effectively collapsed due to the respective governments' inability to maintain the exchange rate of their currencies to the dollar, which had been fixed at artificially high levels. This affected such 'tiger economies' as Thailand, South Korea, Malaisia and Indonesia. The Asian collapse undermined investor confidence in emerging markets in general, and combined with the low oil prices, had a dramatic impact on the Russian economy several months later. August 1998 saw the collapse of the ruble and of the Russian stock market, and the 'Asian flu' was spreading further, to Latin America. The financial instruments in LTCM's portfolio, derivatives linked to emerging market securities, were suddenly worth a lot less than before.

[141] By a 'transient' inefficiency, I mean one that has a short-term nature and is not grounded in the economic fundamentals.

How did the devaluation of, say, the Malaysian ringgit cause LTCM's losses? Consider a ringgit-denominated, fixed-rate bond. When the ringgit loses half its value in relation to the US dollar, all interest payments on the bond, while constant in ringgit terms, become worth only half their previous value in dollar terms. In the case of a 'fixed-term' bond, the same also applies to the bond's 'redemption value', i.e., the sum which is repaid to the bond holder on 'maturity' (the end of the bond's term). And since the value of a bond is nothing other than the future-discounted sum of interest payments on it and, for fixed-term bonds, of the final redemption value, the bond has effectively halved in value. Moreover, since the currency devaluation resulted from underlying macroeconomic problems, the bond issuer (a company, local authority or even the national government) may not be able to redeem the bond when it matures or even to service the regular interest payments. This possibility further decreases the market value of the bond.

Things are not much rosier if the bond is dollar-denominated. In other words, the issuer pledges to pay the interest and the redemption value in US dollars, regardless of the current dollar exchange rate of the local currency. That pledge is, however, subject to the issuer's actual ability to pay. Take a company or provincial government which receives most of its income in the newly-devalued ringgit. Its ringgit income is no longer sufficient to buy enough dollars for the required repayments. In the case of a commercial issuer, the ability to pay may be further diminished by having to buy materials abroad, which at the new exchange rate has become twice as expensive as before. Finally, the economic difficulty which resulted in the devaluation will have depressed the local market, thus decreasing the company's ringgit revenue. All these factors conspire to reduce the company's ability to fulfil its financial obligations and may even bankrupt it, in which case the bond holder gets back nothing.

Collapse and rescue

As an example of the immediate impact the 'Asian flu' had on LTCM's investors, consider the case of Taiwan's Chinatrust Commercial Bank. According to the bank's own estimates, as of the end of August 1998 its US$50 million stake in LTCM had

shrunk by US$26 million, or 52% of the investment.[142] However, the losses to the fund's direct investors paled into insignificance compared with the losses threatening its lenders and counter-agents. According to some estimates, the total value of the bets in which LTCM was involved, directly or indirectly, represented US$1.25 trillion. Given LTCM's high leverage, i.e. its degree of indebtedness, the fund's ability to service its loans crucially depended on its continued ability to turn out a profit. Once the fund started making losses thus decimating its own capital base, its credit-worthiness plummeted and it found itself unable to make regular loan repayments. Moreover, the erosion of LTCM's capital base meant that it was unable to close its open positions, i.e., fulfil its contracts to its many counter-parties.

Picture a 'future' contract for security x. The counter-party paid for it upfront, in return for the promise of delivery at a later date. Of course, LTCM would not have the security at the time of concluding the contract, intending to buy it on the 'spot' (i.e., immediate-delivery) market later. When the delivery date approached, LTCM would no longer have the funds to buy the security in order to fulfil its contract. The counter-party would be in danger of having paid good money for nothing, resulting in its own financial position being compromised.[143]

Thus, had LTCM been left to its own devices, it would have had to default both on its loans and on its contracts, thus leaving several major financial institutions which were not its investors, billions dollars out of pocket. This would have been unlikely to drive many of them to the wall, but profit margins would have been severely squeezed, and the banks' ability to lend dramatically limited, at least in the short term.[144] More importantly, this would have sent a signal to other financial institutions to cut down on lending, leading to a so-called 'credit crunch' damaging to both the US and the world economies:

---

[142] Hong Kong Standard, 29 September 1998.

[143] Although LTCM generally gambled on a rising market, its portfolio also contained large volumes of bets on individual securities falling in price.

[144] As it happened, such major banks as Germany's Dresdner Bank, Switzerland's UBS and UK's Barclay's Bank, suffered the losses of hundreds of millions dollars.

> Because hedge funds are so dependent on banks for the loans that make their sophisticated financial gambles possible, the collapse of a prominent fund such as LTCM can have an expansive effect across the economy, eroding confidence as it goes (Herbert 1998).

LTCM's collapse was made possible by its investment policy, but its direct cause was a unique combination of adverse market conditions. The same conditions, namely the spreading lack of investor confidence, which caused the collapse, at the same time made its implications potentially much more far-reaching than they would have been at other times. As the waves of risk aversion spread to other market agents, business activity would have been dampened still further. An all-out 'great recession' would have been unlikely, but America's 'Goldilocks economy'[145] might have been significantly damaged. The US Federal Reserve was not going to take that chance. Therefore, the chairman of the Federal Reserve Alan Greenspan, arguably the single most powerful individual in the world economy, put together a rescue package for LTCM, designed to decrease the repercussions of its failure.

Under the package, LTCM's 14 largest creditors injected US$3.75 billion of fresh capital into the fund, thus enabling it to close its open positions and remain in business. In return, they took 90% of the fund's equity, thereby displacing the original investors as LTCM's new owners but, given the complexity of LTCM's portfolio, choosing to keep the existing management. An alternative for them would have been to have the fund wound down, its assets sold off and the proceeds used to partially repay LTCM's loans. Of course, the remaining value of the fund's assets would have covered only a small proportion of its debts. Worse, such a 'fire sale' would have raised much less value than could eventually be recovered if the fund continued trading. Part of the reason why hedge funds manage to produce above-average returns, is that they often trade in obscure or illiquid investments such as Danish mortgages, Cuban sugar futures or even pork bellies. Unlike, e.g., Glaxo Wellcome shares or US Treasury bills, millions of which are traded every day, the market for pork bellies is very shallow, in that no actual trades may occur for a long time. Therefore, if you must sell them in a hurry, chances are you will not get a very good price. In addition, if such a fire-sale happens on a large

[145] It is said that since the early 1990s the US economy has been 'neither too hot, nor too cold', just like the bears' gruel in the Goldilocks tale. This means that economic growth has been neither too strong to cause high inflation, nor too slow to cause high unemployment.

scale, as would have been the case with LTCM, all the other holders of such illiquid investments will panic and try to dump them for the security of US Treasury bills. Consequently, large swathes of the economy will find themselves without adequate funding. As Greenspan said in his deposition to Congress,

> Financial market participants were already unsettled by recent global events. Had the failure of LTCM triggered the seizing up of markets, substantial damage could have been inflicted on many market participants, including some not directly involved with the firm, and could have potentially impaired the economies of many nations, including our own. With credit spreads already elevated and the market prices of risky assets under considerable downward pressure, Federal Reserve officials moved more quickly to proved their good offices to help resolve the affairs of LTCM than would have been the case in more normal times. In effect, the threshold of action was lowered by the knowledge that markets had recently become fragile. Moreover, our sense was that the consequences of a fire sale triggered by cross-default clauses, should LTCM fail on some of its obligations, risked a severe drying up of market liquidity. The plight of LTCM might scarcely have caused a ripple in financial markets or among federal regulators 18 months ago - but in current circumstances it was judged to warrant attention (Greenspan 1998).

This was the reasoning that led the Federal Reserve to propose a bail-out plan, and LTCM's creditors to accept it. The rescue was not without its costs, as the dollar slipped to a 17-month low against the pound and banking shares lost value at all the main stock exchanges. However, these costs might have been even higher without the intervention. Arguably, among the most undesirable consequences of the Federal Reserve's initiative was the signal it sent to firms around the world that the relevant government or central bank may mitigate the cost of future mistakes. The danger, known as 'moral hazard', is that the firms will behave less responsibly than they would have done without such a precedent. Still, the bail-out achieved its stated purpose of calming down the world financial markets and preventing a credit crunch.

Nobel Prize-winning methodology

Having placed the story of LTCM's failure and subsequent rescue in its context, it is now instructive to discuss the methodology based on which the fund was hoping to beat the market. The brains behind the fund were two 1997 Nobel prize winners in

Economics, Robert C. Merton and Myron S. Scholes. The prize was awarded for the work they (and Scholes' collaborator, the late Fischer Black) did in the 1970s.

The problem which Merton, Scholes and Black addressed was that of the valuation of options. An option is a financial instrument that allows the holder to purchase, at some future point (the date of maturity) the underlying security, such as a share or a bond, at a previously agreed price (the strike price)[146]. It is normally expected that the underlying security will have appreciated in value by the time it is exercised compared to the strike price, thus allowing the holder to purchase the security at a discount to the market value and make a profit. By buying the option, the purchaser is paying for this chance of making a profit. Of course, the underlying security may fail to perform up to expectation, in which case its market price on maturity may be lower than the strike price. Rather than buy the security at a premium, the option holder will then simply decline to exercise the option (unlike a future, whose holder is contractually obliged to buy the underlying security at the strike price, be it above or below the market price on maturity). If the option is not exercised, the only loss is the price paid for the option in the first place. Of course, at the time of purchase it is hoped that the security will perform well enough to be worth the initial outlay, but there is always an element of risk that has to be weighed against the potential reward. The work of Merton, Scholes and Black was designed to enable the buyer of an option to calculate its 'fair' price.

On the surface, there does not seem to be much of a problem, especially to a Bayesian. Take the potential purchaser's utility of money function and his prior probability distribution with regards to the performance of the underlying security, and integrate the expected utility of buying the share over the range of its possible maturity prices. Then calculate the present-discounted money value corresponding to this utility level, and deduct from this the strike price of the security. The result is the (subjectively) fair price of the option. If it is greater than the price demanded, buy the option, decline otherwise. As described, the procedure allows the buyer to weigh the probability of monetary gain against the risk of a loss. Clearly, the more risk-averse the individual (and therefore the

[146] Strictly speaking, the financial instrument thus described is a so-called *call* option, as opposed to a *put* option which functions similarly but gives the right to *sell* a security at the agreed price. For simplicity, I will refer to call options as simply options.

'curvier' his utility function), the greater will be the distance between the money equivalent of the expected utility of buying the option and the expected maturity price. This distance (usually dubbed the 'risk premium'), can be used, along with the curvature of his utility function as a measure of the individual's attitude to risk. Unfortunately, neither risk premia nor utility functions are observable or easily, if at all, computable.

Instead, Merton, Scholes and Black suggested incorporating the risk factor directly into the price of the underlying security, by noticing that its maturity price is positively correlated with its present price. The higher the price today, the likelier it is that the maturity price will exceed the strike price, and therefore the more valuable the option is. The press release of the Nobel Committee for economics describes the prize winners' methodology thus:

> Consider a so-called European call option[147] that gives the right to buy one share in a certain firm at a strike price of $50, three months from now. The value of this option obviously depends not only on the strike price, but also on today's stock price: the higher the stock price today, the greater the probability that it will exceed $50 in three months, in which case it pays to exercise the option. As a simple example, let us assume that if the stock price goes up by $2 today, the option goes up by $1. Assume also that an investor owns a number of shares in the firm in question and wants to lower the risk of changes in the stock price. He can actually eliminate that risk completely, by selling (writing) two options for every share that he owns. Since the portfolio thus created is risk-free, the capital he has invested must pay exactly the same return as the risk-free market interest rate on a three-month treasury bill. If this were not the case, arbitrage trading would begin to eliminate the possibility of making a risk-free profit. As the time to maturity approaches, however, and the stock price changes, the relation between the option price and the share price also changes. Therefore, to maintain a risk-free option-stock portfolio, the investor has to make gradual changes in its composition.
>
> One can use this argument, along with some technical assumptions, to write down a partial differential equation. The solution to this equation is precisely the Black-Scholes' formula. Valuation of other derivative securities proceeds along similar lines. Black and Scholes' formula for a European call option can be written as
>
> $$C = SN(d) - Le^{-rt}N(d - \sigma\sqrt{t})$$

[147] A *European* option is exercisable only on the maturity date, unlike an *American* option which is

where the variable d is defined by

$$d = \frac{\ln \frac{S}{L} + (r + \frac{\sigma^2}{2})t}{\sigma\sqrt{t}}$$

According to this formula, the value of the call option C, is given by the difference between the expected share value - the first term on the right-hand side - and the expected cost - the second term - if the option right is exercised at maturity. The formula says that the option value is higher the higher the share price today S, the higher the volatility of the share price (measured by its standard deviation) sigma, the higher the risk-free interest rate r, the longer the time to maturity t, the lower the strike price L, and the higher the probability that the option will be exercised (the probability is evaluated by the normal distribution function N) (The Royal Swedish Academy of Sciences 1997, 2-3).

A crucial advantage of this formula is that it contains only one non-observable parameter (sigma) which, moreover, can be easily estimated from market data. Since the original paper was published in 1973[148], the Black-Scholes-Merton approach was generalised by applying to other types of securities and by allowing non-zero transaction costs, discontinuous price changes, and relaxing other restrictions. Very quickly, the formula gained acceptance and became widely used by traders and investors alike, as well as generating several new classes of financial instruments. As a result, the Nobel Committee points out, the research in question 'facilitated more efficient management of risk in society' (The Royal Swedish Academy of Sciences 1997i, 6).

The bearing on EMH

While the last statement may sound ironic in view of the events that unfolded a year later, how much blame can be attributed to the Black-Scholes-Merton methodology? The answer is, none at all. The formula itself merely indicates what the fair option on a security is, based on the information available at the present moment. It has nothing to say about what direction the interest rates will take, how share price volatility will

---

exercisable at any time up to and including the maturity date.

[148] Black and Scholes (1973).

change, or how the probability of exercising the option may be affected by exogenous factors. All these depend on a host of macro- and micro-economic conditions which may change at any time rendering the original calculations worthless.

What is to blame is not the formula itself (or any other formula Merton and Scholes may have used as managers of LTCM), but the investors' and their economist advisors' excessive optimism in relation to the applicability and robustness of any given set of formulas. Generally, as the investors' conceptual arsenal expands, and as high-profile managers are recruited to run their assets, the limitations inherent in their investing models become increasingly overlooked. Caution goes out of the window, giving place to exaggerated confidence in the extent to which prices and returns can be predicted. Then, inevitably, reality reasserts itself.

BIBLIOGRAPHY

- [Anon.] 1999. 'What is a hedge fund?'. *The Hedge Funds Home Page.* http://www.hedgefunds.net/whatis.htm.

- Abdel-Khalik, R. et al 1994. 'Statement on derivative markets and financial risk'. *Financial Economists Roundtable.* Stanford University, September 26, 1994. http://www.sharpe.stanford.edu/fer94.htm.

- Achinstein, P. 1994. 'Stronger evidence'. *Philosophy of Science* **61**, 325-50.

- Akaike, H. 1973. 'Information theory and an extension of the Maximum Likelihood Principle'. In Petrov, B. N. and Csaki, F. (eds.) 1973. *2nd International Symposium on Information Theory*, 267-81. Budapest: Akademiai Kiado.

- Akaike, H. 1974. 'A new look at the statistical model identification'. *IEEE Transactions on Automatic Control* **19**, 716-23.

- Akaike, H. 1977. 'One entropy maximization principle'. In P. R. Krishniah (ed.) 1977. *Applications of Statistics*, 27-41. Amsterdam: North-Holland.

- Akaike, H. 1985. 'Prediction and Entropy'. In Atkinson, A. C. and Feinberg, S. E. (eds.) 1985. *A Celebration of Statistics,* 1-24. New York: Springer.

- Arrow, K. 1964. 'The role of securities in the optimal allocation of risk-bearing'. *Review of Economic Studies* **31**, 91-96.

- Benenson, F. (manuscript). 'Two dogmas of subjectivism'.

- Billingsley, P. 1979. *Probability and Measure.* New York: Wiley.

- Black, F. and Scholes, M. 1973. "The pricing of options and corporate liabilities". *Journal of Political Economy* **81**, 637-54.

Bibliography

- Blackburn, S. 1984. *Spreading the Word: Groundings in the philosophy of language.* Oxford: Clarendon Press.

- Brown, H. I. 1994. 'Reason, judgement and Bayes's Law'. *Philosophy of Science* **61**, 351-69.

- Calude, C. S. and Chaitin, G. J. 1999. 'Mathematics: randomness everywhere'. *Nature* **400** (6742), 319-20.

- Campbell, J. Y. and Shiller, R. 1988. 'The dividend-price ratio and expectations of future dividends and discount factors'. *Review of Financial Studies* **1**, 195-228.

- Carnap, R. 1947. *Meaning and Necessity.* Chicago: University of Chicago Press.

- Carnap, R. 1950. *Logical Foundations of Probability.* Chicago: University of Chicago Press.

- Carnap, R. 1952. *The Continuum of Inductive Methods.* Chicago: University of Chicago Press.

- Cartwright, N. 1989. *Nature's Capacities and their Measurement.* Oxford: OUP.

- Cartwright, N. 1995. 'Causal structures in econometrics'. In Little, D. (ed.) 1995. *On the Reliability of Economic Models: Essays in the philosophy of economics*, 63-74. Dordrecht: Kluwer.

- Church, A. 1940. 'On the concept of a random sequence'. *Bulletin of the American Mathematical Society* **46**, 130-35.

- Cosmides, L. 1989. 'The logic of social exchange: has natural selection shaped how humans reason? Studies with the Watson selection task'. *Cognition* **31**, 187-276.

- de Finetti, B. 1931. 'Probabilism' (English translation). *Erkenntnis* 31 (1989), 169-223.

Bibliography

- de Finetti, B. 1937. 'Foresight: its logical laws, its subjective sources' (English translation). In Kyburg, H. E. Jr. and Smokler, H. E. (eds.) 1980. *Studies in Subjective Probability*, 53-118. Huntington, New York: Robert E. Krieger.

- de Finetti, B. 1977. 'Probability: beware of falsifications!' (English translation). In Kyburg, H. E. Jr. and Smokler, H. E. (eds.) 1980. *Studies in Subjective Probability*, 193-224. Huntington, New York: Robert E. Krieger.

- DeVito, S. 1997. 'A gruesome problem for the curve-fitting solution'. *BJPS* **48**, 391-96.

- Dorling, J. (manuscript). 'Probability and simplicity are just two sides of the same coin and both are just a matter of counting bits'.

- Fama, E. F. 1970. 'Efficient capital markets: a review of theory and empirical work'. *Journal of Finance* **25** (2), 383-417.

- Fama, E. F. 1976. *Foundations of Finance*. New York: Basic Books.

- Fama, E. F. 1991. 'Efficient capital markets II: a review of theory and empirical work'. *Journal of Finance* **46** (5), 1574-1617.

- Fama, E. F. and French K. R. 1988. 'Dividend yields and expected stock returns'. *Journal of Financial Economics* **22**, 3-25.

- Festa, R. 1996. *Cambiare Opinione. Temi i Problemi di Epistemologia Bayesiana.* Bologna: CLUEB.

- Fisher, R. A. 1970. *Statistical Methods for Research Workers*, 14th edition (first published 1925). Edinburgh: Oliver and Boyd.

- Forster, M. 1995. 'Bayes and bust: simplicity as a problem for a probabilist's approach to confirmation'. *BJPS* **46**, 399-424.

- Forster, M. 1995i. 'The Golfer's Dilemma: a reply to Kukla on curve-fitting'. *BJPS* **46**, 348-60.

Bibliography

- Forster, M. and Sober, E. 1994. 'How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions'. *BJPS* **45**, 1-35.

- Forster, M. and Sober, E. (forthcoming). 'Why likelihood?'.

- Galavotti, M. C. 1989. 'Anti-realism in the philosophy of probability: Bruno de Finetti's subjectivism'. *Erkenntnis* **31**, 239-61.

- Galavotti, M.C. 1992. Probabilità, Induzione, Metodo Statistico. Bologna: CLUEB.

- Gillies, D. 1973. *An Objective Theory of Probability*. London: Methuen.

- Gillies, D. 1991. 'Intersubjective probability and confirmation theory'. *BJPS* **42**, 515-34.

- Gillies, D. 1998. 'Confirmation theory'. In Gabbay, D. M. and Smets, Ph. (eds.) 1998. *Handbook of Defeasible Reasoning and Uncertainty Management Systems, vol. 1*, 135-67. Kluwer Academic Publishers.

- Gillies, D. (forthcoming). 'Propensity and causality'.

- Glymour, C. 1980. *Theory and Evidence. Princeton.* New Jersey: Princeton University Press.

- Good, I. J. *Probability and the Weighing of Evidence.* London: Griffin.

- Good, I. J. 'The paradox of confirmation'. *BJPS* **11**, 63-64.

- Goodman, N. 1854. *Fact, Fiction and Forecast.* London: The Athlone Press.

- Greenspan, A. 1998. 'Private-sector re-financing of the large hedge fund, Long-Term Capital Management.' *Testimony before the Committee on Banking and Financial Services, U.S. House of Representatives*, October 1, 1998. http://www.bog.frb.fed.us/boarddocs/testimony/19981001.htm.

- Hacking, I. 1965. *Logic of Statistical Inference.* Cambridge: CUP.

Bibliography

- Herbert, D. 1998. 'How hedge funds affect you'. *The CNN Financial Network*, October 2, 1998. http://cnnfn.com/quickenonfn/investing/9810/02/1_hedge.

- Hintikka, J. 1966. 'A two-dimensional continuum of inductive methods'. In Hintikka, J. and Suppes, P. (eds.) 1966. *Aspects of Inductive Logic*, 113-32. Amsterdam: North Holland.

- Hoover, K. D. 1990. 'The logic of causal inference'. *Economics and Philosophy* **6**, 207-234.

- Hoover, K. D. 1995. 'Comments on Cartwright and Woodward: causation, estimation, and statistics'. In Little, D., ed.. *On the Reliability of Economic Models: Essays in the philosophy of economics*, 75-90. Dordrecht: Kluwer.

- Horwich, P. 1982. *Probability and Evidence*. Cambridge: CUP.

- Howson, C. 1973. 'Must the logical probability of laws be zero?' *BJPS* **24**, 153-63.

- Howson, C. 1987. 'Popper, prior probabilities and inductive inference'. *BJPS* **38**, 207-24.

- Howson, C. 1988. 'On the consistency of Jeffreys's Simplicity Postulate, and its role in Bayesian inference'. *The Philosophical Quarterly* **38**, 68-83.

- Howson, C. 1997. 'Bayesian rules of updating'. *Erkenntnis* **45**, 195-208.

- Howson, C. and Urbach, P. 1993. *Scientific Reasoning: the Bayesian approach*, 2nd edition. Chicago and La Salle, Illinois: Open Court.

- Howson, C. and Urbach, P. 1994. 'Probability, uncertainty and the practice of statistics'. In Wright, J. and Ayton, P. 1994. *Subjective Probability*, 39-52. Chichester, UK: John Wiley and Sons.

- Jeffrey, R. C. 1983. *The Logic of Decision*, 2nd edition. Chicago: University of Chicago Press.

Bibliography

- Jeffreys, H. 1961. *Theory of Probability*, 3rd edition. Oxford: Clarendon Press.
- Jeffreys, H. and Winch, D. 1921. 'On certain fundamental principles of scientific enquiry'. *Philosophical Magazine* **42**, 269-98.
- Keynes, J. M. 1921. *A Treatise on Probability*. London: Macmillan.
- Kolmogorov, A. N. 1950. *Foundations of the Theory of Probability* (English translation; first published 1933). New York: Chelsea.
- Kolmogorov, A. N. 1965. 'Three approaches to the quantitative definition of information'. *Problemy Peredachi Informatsii* 1, 4-7.
- Kyburg, H. E. Jr. and Smokler, H. E. (eds.) 1980. *Studies in Subjective Probability*. Huntington, New York: Robert E. Krieger.
- Leamer, E. E. 1986. 'Bid-ask spreads for subjective probabilities' in Goel, P. and Zellner, A. (eds.) 1986. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 217-32. Amsterdam: North-Holland.
- LeRoy, S. F. 1989. 'Efficient capital markets and martingales'. *Journal of Economic Literature* **27** (4), 1583-1621.
- LeRoy, S. F. and Porter, R. D. 1981. 'The present value relation: tests based on implied variance bounds'. *Econometrica* **39**, 555-74.
- Lindley, D. V. and El-Sayad, G. M. 1968. 'The Bayesian estimation of a linear functional relationship'. *Journal of the Royal Statistical Society* **30B**, 190-202.
- Little, D. (ed.) 1995. *On the Reliability of Economic Models: Essays in the philosophy of economics*. Dordrecht: Kluwer.
- Lonsdorf, S. A. 1998. *Statement to the Committee on Banking and Financial Services, U.S. House of Representatives*, October 1, 1998. http://www.vanhedge.com/tstimony.html.

Bibliography

- Lukas, R. E. 1978. "Asset prices in an exchange economy". *Econometrica* **46**, 1429-1445.
- Mackie, J. L. 1963. 'The paradox of confirmation'. *BJPS* **38**, 265-77.
- Mackie, J. L. 1980. *Cement of the Universe*. Oxford: Clarendon Press.
- Mises, R. von 1957. *Probability, Statistics and Truth* (2nd English translation; first published 1928). London: George Allen and Unwin.
- Maher, P. 1993. *Betting on Theories*. Cambridge: CUP.
- Niederhofer, V. and Osborn, M. F. M. 1966. 'Market making and reversal on the stock exchange'. *Journal of the American Statistical Association* **61**, 897-916.
- Officer, R. R. 1973. 'The Variability of the Market Factor of the New York Stock Exchange'. *Journal of Business* **46**, 434-453.
- Parikh, A. and Bailey, D. 1990. *Techniques of Economic Analysis with Applications*. New York, London etc.: Harvester Wheatsheaf.
- Popper, K. R. 1957. *The Poverty of Historicism*. London: Routledge & Kegan Paul
- Popper, K. R. 1959. 'The propensity interpretation of probability'. *BJPS* **10**, 25-42.
- Popper, K. R. 1959i. *The Logic of Scientific Discovery*. London: Hutchinson.
- Popper, K. R. 1963. *Conjectures and Refutations*. London: Routledge and Kegan Paul.
- Pratten, C. F. 1993. *The Stock Market*. Cambridge: CUP.
- Ramsey, F. 1926. 'Truth and probability'. In Kyburg, H. E. Jr. and Smokler, H. E. (eds.) 1980. *Studies in Subjective Probability*, 23-52. Huntington, New York: Robert E. Krieger.

Bibliography

- The Royal Swedish Academy of Sciences 1997. 'The Bank of Sweden Prize in economic sciences in memory of Alfred Nobel 1997'. http://www.nobel.se/announcement-97/economy97.html.

- The Royal Swedish Academy of Sciences 1997i. 'Background on the Bank of Sweden Prize in economic sciences in memory of Alfred Nobel 1997'. http://www.nobel.se/announcement-97/ecoback97.html.

- Samuelson, P. 1965. 'Proof that properly anticipated prices fluctuate randomly'. *Industrial Management Review* **6**, 41-49.

- Savage, J. L. 1954. *The Foundations of Statistics*. New York: Wiley.

- Savage, J. L. 1980. 'Elicitation of personal probabilities'. In Kyburg, H. E. Jr. and Smokler, H. E. (eds.) 1980. *Studies in Subjective Probability*, 147-192. Huntington, New York: Robert E. Krieger.

- Urbach, P. 1987. 'Clinical trial and random error'. *New Scientist* **116**, 52-55.

- Van Fraassen, B. C. 1980. *The Scientific Image*. Oxford: Clarendon Press.

- van Fraassen, B. C. 1984. 'Belief and the Will'. *Journal of Philosophy* **81**, 235-256.

- Vinod, H. D. and Ullah, A. 1981. *Recent Advances in Regression Methods*. New York and Basel: Marcel Dekker.

- Wayne, A. 1995. 'Bayesianism and diverse evidence', *Philosophy of Science* **62**, 111-121.