Department of Statistics
University of Stockholm

EPISTEMIC PROBABILITY AND EPISTEMIC WEIGHT

– A means to describe Bayesian posterior distributions?

by

Per Gunnar Berglund

Table of contents

Preface

The background to this essay is my long standing and deep interest in the work of the British economist John Maynard Keynes. Any one who has seriously dealt with matters of economic theory must, sooner or later, delve deeper into the theory of probability.

In his young years, Keynes wrote *Treatise on Probability*[1], which – using his own words – was aimed at sorting out the ”curious relation between ‘probable’ and ‘ought’ ”. It was the dissatisfaction which Keynes’s and the rest of the Bloomsbury group felt with the Cambridge philosopher G.E. Moore’s analysis of ”ethics in relation to conduct” that spawned Keynes’s writing of a treatise on probability.[2] The intuitionist ethics of Moore’s *Principia Ethica* conveys many important insights for the theory of probability, and it is crucial to understand Keynes’s theory of probability.

The aim of this essay is certainly more limited than Keynes’s extremely ambitious work. Rather than analysing ”the relation between probable and ought”, I want to argue what we *ought to mean by ‘probable’*. Nota bene this is a ”normative” aim – it is about what the concept of ”probability” *ought to* denote, not what it actually denotes.

Probability is such a vague and multifaceted concept that a study of its actual usage would result in a veritable snake-pit. At any rate, it would not avail itself to comprehension within the limits set by the format of this essay.

My point of departure is that ”probability” should be an objective concept. It should not denote what we *do believe,* but it should denote what *actually is* and what we *actually know*. This may be helpful to keep at the back of one’s mind when reading my account.

[1] *Vide* Keynes (1921).

[2] Moore (1903), chapter 5 – ”Ethics in Relation to Conduct” in particular. For testimonial evidence on Keynes’s and the Bloomsbury group’s attitude towards Moore, *vide* Keynes (1938).

1. Introduction

This essay bears the title ”Epistemic Probability and Epistemic Weight – Means to Describe Bayesian Posterior Distributions?”. The title, somebody said, sounds nice. But what is it really about? The thing is not as difficult as the words. It should be no surprise that the concept of ”probability” is not entirely unambiguous. It can be given various interpretations, often named ”subjective probability”, ”frequentistic probability”, and the like.

”Epistemic probability” is a concept coined by the British statistician Ian Hacking.[3] The adjective ”epistemic” origins from Aristotle’s *Nichomachian Ethics*. *Episteme* roughly translates to ”true and eternal knowledge”, the acquisition of which is one of the three Aristotelian virtues (the others being *techne* – art skills, from which ”technology” stems, and *phronesis* – modesty[4]). In the Anglo-Saxon philosophy literature, the term ”epistemology” is frequently used to denote the theory of knowledge.

Thus, epistemic probabilities have got to do with our possession of knowledge, or information. Hacking carefully distinguishes between epistemic probabilities and ”aleatory” probabilities. As we should have some idea as to what epistemic means by now, we may proceed to sort out the meaning of ”aleatory”. Surely most of you have heard the familiar quotation ”alea iacta est” – the dice are thrown. ”Alea” means die, and ”aleatory” would, somewhat clumsily, translate to ”dice-wise”. But what exactly do we mean by saying that a probability is ”dice-wise”? Are not all probabilities ”aleatory”?

[3] Hacking (1975).

[4] For an easy-to-read presentation, *vide* Flyvbjerg (1994).

2. Aleatory probabilities

All of us are aware that a well-made die has a tendency to fall as often on any of its six sides. After all, it is through dice games and dice experiments we gain our first insights into the peculiar systematic of randomness. We have all patiently thrown dice in the secondary school maths classes and noted the relative frequency of ”sixes”, only to find that is converges towards one sixth as the number of dice rolls looms large. To this has been added that the outright definition of the concept of ”probability” is the limiting value of a relative frequency, as the number of trials becomes large.

Few doubt that probabilities of this ”aleatory” kind basically expresses a physical quality of nature.[5] The world is not deterministic, or pre-set, but ”randomness” really exists and constitutes a highly palpable factor in our lives. And ”randomness” can, to some extent, be systematised by using the concept of aleatory probability, just because some events have a greater propensity to occur than others.

At closer inspection, ”randomness” or ”aleatory probability” shows to be very difficult, not to say impossible, to define. This is by no means unique for this quality of nature. The quality of ”mass”, for example, is not defined, but merely exemplified by an international kilogram prototype. It is left to our imagination what ”mass” really is. I associate myself with those who take the view that ”aleatory probability” must be one of those intuitively intelligible properties.

I do not believe that any real definition of ”aleatory probability” is possible. The closest we can get is to circumscribe it into other, equally indefinable terms. This is a perfectly normal procedure when it comes to the concept of probability, which is being practised by several ”schools” of probability theory.

In the theory of ”subjective probability”, which does not deal with what we call ”aleatory probabilities”, but with something else, the concept ”probability” usually denotes the subjectively held ”degree of belief”.[6]

In the theory of ”logical probability” the concept ”probability” usually denotes a relation between propositions, of the kind »If *p*, then it is probable to the degree α that *q*.«, where *p* and *q* denote simple propositions. One assumes that such a relation always exists between every pair of propositions, and that the probability mirrors the ”rational degree of belief”, i.e. that α is the degree of belief which a creature endowed with perfect logical intuition would

[5] One of those who claim that probability does *not* exist is Bruno de Finetti. *Vide* de Finetti (1990), introduction.
[6] *Vide* e.g. de Finetti (1972), Savage (1954), Ramsey (1926, 1928).

hold before *q*, provided that *p*.[7] It should be pointed out that the concept of ”logical probability” does not refer to the same phenomenon as our concept of ”aleatory probability”.

Thus, what we reasonably can do is to circumscribe our concept of ”aleatory probability” into terms that are intuitively comprehensible. First of all, we must make clear that aleatory probabilities are not states of mind, but rather qualities of the surrounding world, qualities of nature. Therefore, aleatory probabilities have no immediate connection to the subjectively held ”degree of belief”.

On the contrary, aleatory probabilities would correspond to something like the ”rational degree of belief” under certain idealised assumptions. By this I do not mean that aleatory probabilities are ”logical” in the sense that they express relations between propositions. It is rational to hold a degree of belief that accords with the aleatory probability, but only under the assumption that one has full knowledge about the true size of that aleatory probability. The determining factor is the possession of knowledge or information. If one does not possess full knowledge about the true size of the aleatory probability, then it is not certain that a rational usage of the limited of the limited knowledge one actually has will lead to a ”rational degree of belief” that accords with the aleatory probability.

We must carefully distinguish between true properties of the world, and our knowledge about those properties. The mere case that an event is aleatorily probable of the degree Π, that does not necessarily imply that we know it is. But to be able to discuss the magnitude of an aleatory probability, we must first suppose that there is such a thing as ”aleatory probabilities”. For this reason, the question what we know about an aleatory probability subdivides into two parts. The first part is the question how we can know whether there exists such a thing as aleatory probability in nature. If we should answer that question in the negative, we would render the second part – the question of the magnitude of that aleatory probability – meaningless. For this reason it is absolutely necessary to presuppose that aleatory probabilities exist, if we are to discuss their magnitudes.

[7] *Vide* e.g. Keynes (1921), Carnap (1950).

3. Metaphysics and morality

The question whether aleatory probabilities exist is basically metaphysical. It is about the nature of the world, whether we live in a ”stochastic” world (in which aleatory probabilities exist) or in a ”deterministic” world (in which they do not exist). Metaphysical queries, one would think, do not belong to this kind of work.

The thing is we just cannot let the question pass, but we must make clear, or at least roughly clear, what kind of decision we make when we take on a particular metaphysical stance. My own view is that this is chiefly a *moral* decision.

For centuries, moral philosophers and theologians have discussed the question whether or not man has a free will. Obviously the notion of a free will is intimately connected to the notion of moral responsibility for our actions and deeds. For if the world would be strictly predetermined by faith, so that we never ever really would have a choice, we could neither be blamed for our misdeeds, nor praised for our kind actions.

Good and evil are merely fictions in a deterministic world, and the very thought of a deterministic world is so absurd that we must dismiss it on the sole ground of its consequences to morality. The only practicable stance is the notion that the world is not deterministic, and that we ourselves – at least to some extent – can change the order of things by making our own decisions and pursue actions thereafter. In this respect, man is an ”image of God”, as we can and do constitute what the British economist and philosopher G.L.S Shackle termed ”absolute origins”.[8] Absolute origins to chains of events in the world, scenarios conditioned by and affected by our voluntary decisions.

But if the future is not deterministic, or ruled by fate, what is it then? Even if ”randomness” should not exist as a phenomenon in nature, the future development of the world would still not avail itself to exact prediction, since it will be affected by decisions made by human beings. Free will, our passing fancies, whims and caprice, our considered doings, will render that part of the world, which lies within our powers to affect, unpredictable, at least to some extent. This would still be so, even if the *consequences* of our decisions at a certain point in time, once these decisions were made, would be (at least in principle) possible to determine exactly. This is so because of the possibility of new decisions in the future changing the scenario implied by decisions made earlier.

But it is not even certain that a particular array of decisions will lead to determinable consequences. The really huge decision, if we may talk about anything like that, was God’s

[8] *Vide* Shackle (1974).

decision to set the world off. Our perception of nature is of that kind. Some power – let us call it God – once set the world machinery in motion, and rigged its development over time by determining the natural laws. Ever since, says our mythology, God has not interfered with the internal affairs of the world, but He has let the whole thing operate according to the initially stated natural laws.

The deterministic perception of nature, which is usually associated with the French 19:th century mathematician Marquis Pierre-Simon de Laplace, implies the very thought that Laplace formulated.[9] A creature of supreme intelligence would, according to this train of thought, if it had access to all correct natural laws, and the locations and moments of all particles in the universe at a specific point in time, be able to compute the development of the universe in every small detail. Today, we would associate this ”supreme intellect” with some kind of super-computer. The modern machine era association has turned the ”creature” to a machine – the Laplace Machine.

Such a deterministic notion of the universe was the fruit of the success of Newtonian mechanics, and this *Weltanschau* dominated science way into our century. Even such a celebrity as Albert Einstein embraced it – ”God does not play dice”, he said. In perfect accordance, Einstein’s theory of relativity is a completely deterministic theory, which – in its present form – is incompatible with the later developed and non-deterministic quantum mechanics.[10]

Quantum mechanics meant a breach with the deterministic perception of nature. The metaphysical moral of quantum mechanics is that even if the natural laws are given by God once and for all, these laws do not exactly determine what is going to happen, but they are merely regularities in what is going to happen. When the world is viewed with the eyes of quantum mechanics, the access to a Laplace Machine no longer helps. No intelligence in the world can predict exactly what is going to happen in the future, even if it has full knowledge of where the world stands at the moment. The quantum mechanical world is genuinely ”stochastic” in the sense that all future events *are* more or less probable – meaning that they have different propensities to occur within a limited time-period. What a Laplace Machine would be able to do, is to compute different future scenarios, each of which has a particular probability, a specific propensity to occur. But not even the Laplace Machine will be able to say which of these scenarios will actually occur.

[9] Laplace is also, paradoxically enough, known for his works on probability theory, in particular Laplace (1814).

[10] For a brilliant popular presentation of the incompatibility of these theories, *vide* Hawking (1989). An easily comprehensible presentation of the history of particle physics from the antiques to our time is Bergström and Forsling (1992).

It is this kind of *Weltanschau* that is associated with the concept of ”aleatory probability”. The aleatory probability of a particular future event does not denote our subjectively held degree of belief, but the rational degree of belief of the Laplace Machine. Since the Laplace Machine know all natural laws, it follows that the rational degree of belief of the Laplace Machine is exactly equal to the real propensity for that particular event to occur. Thus, the best description of ”aleatory probability” is given by expressions like ”the propensity to occur”.

A matter of concern in the stochastic *Weltanschau* is to make clear what makes some events happen and others not. It may be frustrating to be forced to succumb before this query and admit to ourselves that we simply *cannot* determine that in every single case. Still this is what we have to do if we embrace a non-deterministic *Weltanschau*. It is not at all certain that the most probable scenario will be realised. For that reason, the non-determinist mythology must contain a measure of ”events' mystique”.

In the end, ”chance” will determine the non-deterministic system, and the ”will of God” lies near at hand – what remains of it within the frames of the stochastic natural laws, that is – as the unfathomable factor determining what is actually going to happen, and which thereby determines if and when a merely probable event will become a fact, or whether it will not occur at all, and thus form what I have chosen to call a nullity or a non-event.

A *Weltanschau*, and mythology, of this kind comprises the possibility that man has a free will and thus can constitute an absolute origin. In that way, the mythology is compatible with the Biblical thought of man as an image of God, with an unfathomable capacity to generate events, or nullities, within the frame set by the ”Blind Watchmaker” at Genesis. That mythology would be unthinkable in a universe of the kind that Laplace and Einstein and seemed to have conceived.

Strangely, the non-determinist perception of nature thus constitutes a presupposition for our own free will, and hence for the existence of good and evil. The existence of aleatory probabilities is maybe not liable to proof, but the *notion* of their existence is a *moral* necessity. Our choice of mythology must be adapted to man's needs to be a ”moral creature”,[11] and any mythology which denies us the possibility of being moral creatures stand in contradiction to our needs and our nature. The notion of a non-determinist, aleatorily probable world provides for these needs and thus constitutes a good mythology. A good mythology should of course be preferred to, and chosen before, a bad one.

[11] A *Zoon Politikon, vide* Aristotle's *Ethics.*

Whether or not a phenomenon like aleatory probability exists is, to repeat, a metaphysical question. Metaphysical propositions cannot be proved, but only enter axiomatically into a system. I have argued that in our context, the choice of metaphysical axioms is a moral question. The question is therefore wrongly formulated. This is not about whether aleatory probabilities really exist or not, but whether aleatory probabilities *ought to* exist or not. My opinion is that they ought to exist, and according to this view the axiom of future events being aleatorily probable is simply held true. So far the question of existence.

When we have decided to postulate that aleatory probabilities exist, and that all future events in the world are aleatorily probable, the question remains *how probable* those events are. But before we proceed to discuss the magnitudes of aleatory probabilities, we should take some time to reflect upon the very concept of an ”event”.

4. Fundamental concepts in the theory of probability

The universe possesses *extent* in time and space. By setting time and space limits, we may subdivide the whole world into *partial universums*, or *event spaces*. Event spaces are delimited in a suitable way according to the phenomena we want to study. Examples of event spaces are ”Sweden 1994” or ”the LEP accelerator at Stanford 13.05 hours 3 September 1987”, or some similar delimitation. When we talk about the ”world” we will mean such a suitably chosen event space.

Apart from extent in time and space, the world possesses an array of *properties* which signify various places in time and space. These properties, who describe what is the case in space and time, constitute *states of the world*. The world is in a particular, actual state at any historic point in time, and aleatorily probable states at any future point in time.

A complete listing of all properties of the world can seldom or never be carried out. In practice, we limit ourselves to listing the properties who are *relevant* to the problem at hand. Other *irrelevant* properties we leave out of the listing. The states of the world are thus partitioned in two categories, the first of which contains the relevant properties, and the second of which holds the irrelevant properties. When talking about states of the world in the following digression, we will refer to the list of relevant properties.

What is there to decide whether a property is relevant or not? The relevance must be judged from its effects on the object under investigation – the aleatory probability we are seeking. Properties of the world who do not affect, or negligibly affect, the object under scrutiny can safely be bypassed. The sifting of circumstances must be done from a judgement of what is reasonable, founded in our experience. The larger the precision we wish in our studies of a particular aleatory probability, the more carefully compiled and the more extensive must be the list of relevant properties

Events are defined by the presence (or the absence) or a subset of the world’s properties at a certain point in time and a certain extent in space which we shall call the *premises of the event*. The premises of the event is determined by the amount of time and space *occupied* by the event, i.e. the minimum necessary extent in time and space required to ”lodge” the event.

An event *a* is said to *occur* if (the properties) *x* are the case in [*a: r*, *t*] (co-ordinates for the premises of *a*). If *x* is not the case in [*a: r, t*] , we say that *the complementary event a’* to *a* occurs. It is true that [*a: r, t*] = [*a’: r, t*], i.e. that *a* and *a’* share the same premises. Let *y* denote all other relevant properties of the world in [*a: r, t*]. Thus, in [*a: r, t*] it is either true that ”*x* and *y*” is the case (*a* occurs), or that ”not-*x* and *y*” is the case (*a’* occurs).

As an event expresses a state of the world, it is true that for any future event a of the type A, there is a number Π – the aleatory probability of a – such that

(4.1.) $0 < \Pi(a) < 1$;
(4.2.) it is true that $\Pi(a) + \Pi(a') = 1$;
(4.3.) for the mutually exclusive events $a, b, c, \ldots$ in Ω, it is true that
$\Pi(a \cup b \cup c \cup \ldots) = \Pi(a) + \Pi(b) + \Pi(c) + \ldots$.[12]

Let us now consider a number of event premises in the world. Let us assume for all these premises that ”only x and y” is the case, or that ”only not-x and y” is the case. Thus, the events in each of these premises are equal in all their relevant properties. The only thing separating the events is their position in time and/or space. We then say that the events form a *kind of event* $A = [a_1, a_2, \ldots, a_m]$, with the corresponding *complementary kind of event* $A' = [a'_1, a'_2, \ldots, a'_m]$.

When a kind of event A has been defined, it must be true that a number of event premises exist in event space Ω in which events of the type A are possible. These event spaces we shall call the *A-premises* in event space Ω. The number or A-premises in Ω we shall call the *population m*, the number of which states the highest possible number of events type A in Ω.

Let Ω_A denote the A-premises in Ω, and Ω'_A the remaining part of that event space, or the *surroundings* of Ω_A. If $\Omega'_A = \varnothing$, so that Ω_A lacks surroundings, we shall say that Ω_A is *exhaustive*. When the state of the surroundings Ω'_A is held constant, it holds true, regarding future events type $A = [a_1, a_2, \ldots, a_n]$ in Ω , that

(4.4.) $\Pi(A) = \Pi(a_1) = \Pi(a_2) = \ldots = \Pi(a_m)$.

The whole of this presentation surely appears both abstract and complicated. The reason for its being rather complicated is that an event must be defined as a state at a particular time and place. Since all events (except for the complementary event) cannot occur at that particular time and place, or in those premises as we say, it is necessary to define a type of event from the notion that events are to be alike in all respects but their time and/or place of occurrence.

The reason why we define types of events is that we want to specify the conditions under which we know that *aleatory equiprobability* prevails for all events in the population. But the mere definition of event types then shows to be insufficient, except for the case when the A-

[12] This is Kolmogorov’s axioms. The reference is Kolmogorov (1933).

premises exhaust the entire event space. If not, there will be ”gaps” between the *A*-premises, and in these gaps, other properties may occur, properties which must not be changing if we are to be certain that aleatory equiprobability prevails.

5. Quantitative probability and relative frequencies

Why, then, are we interested in discerning the conditions for aleatory *equi*probability to prevail for a series of future events? The reason for that is that we want to find a method by which to measure *the magnitude of an aleatory probability*. The only way, as far as I can see, to measure aleatory probabilities, is by first making sure that ”laboratory-type” experimental conditions are at hand, i.e. than the surrounding factors – the environment of the experiment – remain unchanged, and that the trials we perform have equally large probabilities to turn out ”favourably”.

By a *favourable* outcome of a trial we mean that an event of type A occurs. If a complementary event of type A' occurs we say that the trial comes out *unfavourably*. The number of favourable outcomes we denote by #(A) ; the number of unfavourable by #(A') .

When aleatory equiprobability prevails, we can show – by the central limit theorem – that the relative frequency of favourable outcomes converge towards a particular determinate proportion between zero and unity, as the number of trials (the *sample n*) becomes ”large”, by which is understood that it approaches the number of ”possible” outcomes, m. This particular proportion is the aleatory probability for a favourable outcome in each individual trial. Thus, it is true, for all future events of type A in Ω, and under the assumption that the state of the surroundings Ω'_A is kept unchanged, that

(5.1.) $\lim_{n \to m} [\#(A)/n] = \Pi(A) = \Pi(a_1) = \Pi(a_2) = ... = \Pi(a_m)$,

and, hence, that

(5.2.) $\lim_{n \to m} [\#(A')/n] = \Pi(A') = \Pi(a'_1) = \Pi(a'_2) = ... = \Pi(a'_m) =$

$= 1 - \Pi(A) = 1 - \Pi(a_1) = 1 - \Pi(a_2) = ... = 1 - \Pi(a_m)$.

It is inadequate to take the step from throwing dice, or pursuing some similar kind of laboratory experiment, and find that the relative frequency *converges* towards a particular value, to *define* the concept of aleatory probability on the basis of a converging relative frequency. If the aleatory probability that a particular kind of event will occur, say, that the die we are holding in our hand, shows a ”six” when thrown, remains the same from trial to trial, then the relative frequency will converge toward this very probability. But that does not validate the opposite – that a converging relative frequency necessarily determines the aleatory probability for a certain kind of event to occur. Let me take a simple example to illuminate this.

Suppose we have an urn containing a large number of black and white balls. The proportions are unknown. With the aid of a mechanical device we draw balls from the urn, without replacement. How large is the aleatory probability of drawing a white ball?

The first thing that springs to mind is of course to draw a decently large sample in order to see how the relative frequency of white balls develops. But that is not enough. What if the balls are of different size, and our mechanical device more often draws large than small balls? And suppose the proportions of white balls is larger among the large balls than among the small? In that case we are no longer dealing with a ”random” sample, and the relative frequency of white balls is likely to diminish the more balls we draw. This is so, because the large balls tend to be drawn first and they are ”whiter” than the small balls that are likely to be drawn later in the sample series.

It is true that there is a limiting value for the proportion of white balls, the knowledge of which we do not get until we have emptied the whole of the urn. But we dare not make any inferences as long as we do not know beforehand that our sample is ”random”, i.e. that all balls are chosen with an aleatory equiprobability.

To ensure that a sample series will give us a fair indication of the population’s composition, we must *first* make sure that the sample draws are made with aleatory equiprobability. In other words, the concept of aleatory probability enters at an earlier stage then the sample series as such.

The example shows that when we are talking about a limited (finite) population size, the mere fact that a certain proportion of ”white balls” (or whatever we are sampling) exists in the population is not sufficient for us to draw the conclusion that the aleatory probability of drawing a ”white ball” is equal to that proportion.

If we, on the contrary, would draw our sample from the urn with replacement, things turn out differently. Then the varying sizes of the balls no longer prevents us from drawing inferences from the sample to the population. This case is interesting for the reason that we have to do with a population that is finite in a way – it consists of the limited number of balls in the urn. But in another way it is unlimited, since we can make as many draws as we like from that population.

When speaking of the ”population”, we refer to the maximum number of trials, which in this case would correspond to the maximum number of balls in the sample. Thus, the population is infinite. Due to the fact that we are dealing with the same balls all the time, which are steadily replaced to that urn, we ensure ourselves that the infinite population has stable properties. The proportion of white and black, of large and small balls is not changing over time. To illuminate the importance of this stability, I would like to give another example.

Assume that we are studying the proportion of deaths in cancer, say, within ten years after the cancer has been diagnosed. Since the number of cancer cases up to now has been very large, and the number of future cancer cases can be said to be unlimited (at the very least we do not know how large it will be), we may safely suppose that the population is unlimited when we include in it both historic and future cancer cases. Historically the proportion of deaths in cancer has been sinking constantly. I do not know the true figure, but let us assume, for the sake of the example, that it has gone down from 90 percent in the year 1900 to about 30 percent today.

The mere fact that the proportion has been sinking gradually over time tells us that there has been no intertemporally stable aleatory probability. The aleatory probability in question, let us denote it by $\Pi(A)$, has obviously gone down over time. For every specific point in time there is a particular, unique value to $\Pi(A)$. This unique value depends on the development level of medical sciences, the access to adequate health services, hygienic conditions, etc. Let us denote this by introducing a time index subscript $\Pi(A_t)$. The space must of course be specified too. $\Pi(A)$ is likely to differ radically between Sweden and Uganda, for example. For this reason a space indexing ought also to be entered, yielding $\Pi(A_{r,\,t})$.

It is characteristic that we are no longer dealing with a population with stable properties, at least not when considering an extended period of time. The population is unlimited, though, which separates this example from the first urn example above. Since the population does not have stable properties, an investigation of $\Pi(A_{r,\,t})$ must be limited to such a short time span that we may assume that the factors influencing $\Pi(A_r)$ remain more or less unchanged. What we are looking for is what the economists call *ceteris paribus* conditions – that ”everything else remains the same”. For it is only when the relevant environmental conditions are unchanged that the relative frequency may be used to *quantitatively estimate* aleatory probabilities.

It is the very fact that ”the relevant conditions remain unchanged” that enable us to make inferences when drawing with replacement from our urn in the prior example. In that case the population had unchanging properties over time, despite its being unlimited. In the cancer example the population’s properties were changing over time, which dashed our hopes to use relative frequencies to estimate probabilities.

In short, the conclusions of this somewhat tedious argument are the following:

Firstly, the concept of aleatory probability cannot be defined. It is an intuitively comprehensible concept which denotes the ”propensity to occur” of a particular event. The existence of aleatory probabilities, and an aleatorily probable world, is a moral necessity.

Secondly, aleatory probabilities cannot be quantitatively estimated unless we *first* make sure that conditions prevail that guarantee a series of *aleatorily equiprobable* events. This requires that (1) the relevant properties of the population and the surroundings are stable, and (2) that the events make up what was defined above as a type of event, i.e. that they are equal with the exception for their location in time and/or space.

The first condition – that the relevant properties of the population and the surroundings remain unchanged – reminds of what is usually called the *ceteris paribus* assumption in economics (that ”everything else remains the same”), which ensures us ”laboratory type” conditions where influencing factors may be isolated. This condition is indispensable for us to be able to estimate aleatory probabilities.

The second condition – aleatory equiprobability – is needed because a population can contain several different properties (black–white/large–small ball), where a correlation exists between the occurrence of the different properties. Therefore it is necessary to define a type of event in such a way that no relevant properties may separate the events from one another. In our urn example this means that white balls may only be white – they must not be big or small besides that. In that case ”white balls” do not constitute a specific type of event.

The two conditions define what Ian Hacking is going for when he speaks of a *Chance Set-up*,[13] or an ”aleatory rigging”. An aleatory rigging is a prerequisite for the estimation of aleatory probabilities, and provided this prerequisite is fulfilled, the aleatory probability will coincide with the limiting value of the relative frequency when the number of trials becomes large.

An aleatory rigging fills the function to guarantee that events with the properties we seek are equiprobable. If we know that the events are equiprobable, it follows with certainty that the relative frequency will converge towards a specific value between zero and unity. This is true regardless of the population being limited or not.

[13] Hacking (1965), chapter 2. References to Cournot, Venn, von Mises, Popper and more are found throughout Hacking (1965).

The relative frequency does not fill a *defining* function, but only a *quantifying* function. Aleatory probabilities cannot be defined, but they avail themselves to quantification – but only under the particular forms which constitute an aleatory rigging. Since aleatory *equi*probability is a prerequisite for the relative frequency to converge, aleatory probability must be a deeper, or prior, concept than the relative frequency as such.

If the limiting value of the relative frequency would *define* the concept of aleatory probability, we would first have to be able to define an aleatory rigging without making use of the concept of probability in that definition. For if we use ”aleatorily equiprobable” in the definition of the events generated in an aleatory rigging, and then use the aleatory rigging to define the concept of aleatory probability – then we are reasoning in a logical circle.

What we have above is to *postulate* two things: (1) that all future events have an aleatory probability, and (2) that all similar events, which constitute a type of events, have *equal* aleatory probabilities provided that the relevant properties of the surroundings remain unchanged. Postulate (2) encapsulates the very notion of ”aleatory equiprobability”, which means that the converging relative frequency is a *consequence* of the aleatory rigging, not a prerequisite for it.

6. Kolmogorov’s axioms and different types of probabilities

Most treatises on the theory of probability[14] put Kolmogorov’s three axioms at the forefront. They are usually described as a mathematical ”least common denominator” to an array of different philosophical interpretations of the concept of probability. In that way, it is argued, the calculus of probability can be taken as common to all various interpretations. In that way, the mathematical side of the theory of probability is separated from the philosophical side.

Still, Kolmogorov’s axiom’s are usually presented in terms of *events*. But probabilities need not refer to events. In the theories of Keynes and Carnap, probabilities refer to *relations between propositions*. The Keynes–Carnap kind of probabilities are often called *logical probabilities*.

One of the earliest explicit formulations of this principle, we find in Keynes’s *Treatise on Probability*.[15] Keynes’s idea is that probabilities express ”relations of partial implication [RPI]” between propositions. A RPI is a ”softer” variety of logical implication. In ordinary, demonstrative logic, the conclusions necessarily follow from the premises. The logic is two-valued, and the operators either yield the value ”true” or ”false”, zero or unity.

Keynes conceived that a corresponding kind of logic could be constructed for positions in between true and false, so that premises could support a conclusion partially. The better the support for the conclusion, the higher the probability. The lowest value for the probability is nil, which means that the premises make impossible the conclusion. The highest value is unity, which means that the premises make necessary the conclusion.

There are interesting connections between Keynes’s concept of probability and what we have chosen to call aleatory probabilities in the foregoing. Keynes emphasises that the construction of a situation of *equiprobability* is required to make possible quantitative measurement of his logical probabilities, and also for the ordinary theory of probability calculus to be applicable. Even if Keynes’s theory was formulated before Kolmogorov’s axioms, there can be little doubt that Keynes meant that the quantitatively measurable probabilities would obey these axioms.[16]

The parallel to our discussion of aleatory probability, and the need for equiprobability to make them quantifiable, should be obvious. Nevertheless the two concepts are entirely separate. Aleatory probabilities are concerned with events in the world, and their propensities to occur.

[14] A good representative familiar to Swedish students is the presentation in Blom (1980), 23–24.
[15] Keynes (1921).
[16] *Vide* Keynes (1921), particularly chapters 4 and 5.

Keynes’s probabilities are, to repeat, logical relations between proposition, for which there is basically no need at all to assume any connection to the properties of the ”world”.

Another category of probabilities is the ”degree of belief” of the theory of subjective probability. What a subject believes, or disbelieves, does not either necessarily have anything to do with the properties of the ”world”. Nor needs a belief pertain to events of the world. For the time being, I will not enter into a discussion of subjective probabilities, but only conclude that such probabilities are also, with the aid of certain methods of measurement, liable to quantification by numbers between zero and unity. Such probabilities may also be assumed to obey the Kolmogorovian axioms.

The conclusion here is that it is true that the Kolmogorovian axioms can be said to constitute a mathematical ”least common denominator” for the calculus of probability. And the calculus of probability is the same in most approaches to the concept of probability. But it is incorrect to formulate Kolmogorov’s axioms in terms of events, since there are interpretations of the concept of probability that do not involve events.

It is also important to remember that Kolmogorov’s axioms only define a particular kind of magnitude with certain mathematical properties. This purely mathematical magnitude we shall call a *Kolmogorov weight*. Kolmogorov weights are functions of the subset Θ_i , $i = 1, ..., n$ of some suitable set Ω (Ω does not necessarily denote what we have called an event space above).

In the theories of probability, Θ_i usually denotes *propositions* (subjective and logical probability theory) or *events* (subjective and aleatory probability theory). But in principle, Θ_i may denote anything subject to weighting by a linear scheme of weights, and where the weights sum up to unity in Ω, i.e. weighting by Kolmogorov weights.

Kolmogorov’s three axioms may generally be formulated in the following way:

(6.1.) For every Θ_i in Ω , there is a real number Q – *the Kolmogorov weight of* Θ_i – such that $0 \leq Q(\Theta_i) \leq 1$;

(6.2.) it is true that $Q(\Theta_i) + Q(\Theta'_i) = 1$, where Θ'_i denotes all other Θ_j in Ω , $i \neq j$;

(6.3.) for the mutually exclusive subsets $\Theta_1, \Theta_2, \Theta_3, ...$ in Ω, it is true that $Q(\Theta_1 \cup \Theta_2 \cup \Theta_3 \cup ...) = Q(\Theta_1) + Q(\Theta_2) + Q(\Theta_3) + ...$.

The point of expressing Kolmogorov weights on this level of generality is that we are now free to apply them to a number of different theories of probability, whether concerned with events or propositions, or something completely different. In that way, the Kolmogorov weights really become the ”least common denominator” of probability calculus which we usefully avail ourselves of.

But there is more to it. For Kolmogorov weights need not at all be interpreted in terms of probabilities. I many other situation, the construction of weighted averages for instance, we use Kolmogorov weights without even thinking in terms of probability. Thus, an asymmetry exists between probabilities and Kolmogorov weights: *A quantitative probability is always a Kolmogorov weight, but a Kolmogorov weight need not express a probability*. Quantitative probabilities thus express a special case of Kolmogorov weights, namely the case where these weights express the magnitudes of probabilities.

This logical relation of implication between the magnitude of probabilities and Kolmogorov weights is of great importance to grasp when we in due time will pass on to discuss epistemic probabilities. But before we do, we will first discuss the problem how to choose models for aleatory riggings.

7. Stochastic variables, densities

Assume a function X from Ω to R^1. X we shall call a *stochastic variable*. X is *discrete* if it can take on a finite or countable infinite number of values. X is *continuous* if the first derivative of the distribution function exists throughout the entire definitional set. We define the *distribution function* of X as

(7.1.) $F(X) = Q(X \leq x) \quad ; \quad -\infty < x < \infty$

the *frequency function*

(7.2.) $f(X) = Q(X = x) \quad ; \quad -\infty < x < \infty$

in the discrete case; and the *density function*

(7.3.) $q(X) = \mathrm{d}F(X)/\mathrm{d}X$

in the continuous case. $q(x)$ we shall call the *Kolmogorov density* in x.

Remember that a quantitative probability is always a Kolmogorov density, but not the reverse. When a Kolmogorov density is complemented with a probability interpretation we call it a *probability density*. Thus, an aleatory probability density is a quantitative probability density with an aleatory probability interpretation, etc. It is also true that a quantitative probability density is always a Kolmogorov density, but not the reverse.

We will use lower-case characters throughout to denote densities. Kolmogorov weights are denoted by a capital Q, and Kolmogorov densities by lower-case q; aleatory probabilities are denoted by a capital Π, and aleatory densities by lower-case π, etc.

8. Aleatory rigging and the choice of models

An aleatory rigging means that the type of event A in Ω , which we are dealing with, can be described by a Bernoulli-type stochastic variable. Let X denote that variable. It is true that

(8.1.) $X = 1$ if a occurs
$X = 0$ if a' occurs

The series of events of type A (and A') may be described by a vector of the numbers zero and unity, arranged in the temporal and/or spatial sequence that a and a' occur.

Since all events are equiprobable in an aleatory rigging, it follows that the series of Bernoulli trials either is binomially or hypergeometrically distributed. If the population is unlimited, the former is true; if it is limited (finite), the latter is true.

Limited populations cause difficulties in certain cases. The cases I have in mind are when we do not know *how large* the population is. It is easy to conceive a situation in which m must be limited, but where we do not know the size of m. These cases are solvable if we can find a probability distribution for m. But this brings on an unnecessary complication in our context, and therefore we will not analyse this case further.

It should be pointed out that a finite population must be rather small (say, $m < 50$) for the use of a hypergeometic distribution in numerical computations to make any significant difference, compared to computations using the binomial distribution. The former converge to the latter when m grows large. In practice, thus, we apply the binomial distribution to large populations, and the hypergeometric to small, where the limit between ”small” and ”large” must be set to match the degree of precision we require in computation.

Thus, the choice of models is very simple in an aleatory rigging. Small population → hypergeometric distribution; large population → binomial distribution. Let $Y = [X_1, X_2, ..., X_n]$ denote the variable created by the repeated Bernoulli trials. It is true that

(8.2.) $Y \in hyp(m, n, \Pi)$

(8.3.) $\lim_{m\to\infty} hyp(m, n, \Pi) = bin(n, \Pi)$

What we are to discuss in the following, and what epistemic probability and epistemic weight is about, is the *estimation* of one of these model parameters, namely the aleatory probability Π.

The problem with aleatory probabilities is that even if we know that they exist (a morally necessary assumption), in many cases we can still not know *how large* they are. As I mentioned initially, the epistemic side of the probability theory deals with our *possession of knowledge*.

To really know the true value of Π , it is necessary to possess full knowledge of the entire population. In some simple cases (like the urn, for example), it is easy to find out the relevant proportion in the population as a whole. But usually we must make do with knowing only a part of the population, and then we must make *inferences* from this part to the whole.

To an omniscient creature, a Laplace Machine knowing all natural laws, such an inference problem never occurs. The Laplace Machine always know the entire population; for such a creature, there is no difference between the factual and its knowledge thereof. The inference problem does not exist for such a creature – the theory of statistical inferences is there for us common mortals, who often must make do with incomplete knowledge and partial information.

9. The inference problem

Our inference problem consists in our possessing an aleatory rigging of an event type *A* in Ω , so that we know that an aleatory probability Π(*A*) exists and that Π(*A*) has a particular numerical value between zero and unity. But we do not know exactly what this value is. We may have a rather good hunch about the value, by our knowledge of parts of the contents of Ω. To be able to systematise this, we must first define some more concepts.

As soon as an event has occurred, its passes from being aleatorily probable to being a *fact*. Thus, history is made of facts, and the future of aleatorily probable events. Facts are either *known* or *unknown*. Either we know that the event *a* was the case, or we don't.

We cannot know for sure whether future events really are going to occur, or if unknown facts really did occur. But for both these phenomena, we can – provided they can be regarded as spawned by an aleatory rigging – make inferences which characterise our knowledge position with regard to the type of event at hand.

Assume, for instance, that we are dealing with historic facts of the type *A* (and *A*') in an event space Ω. Some of all these facts we know, while another part of them is unknown to us. If we can assume that the generation of these facts *A* and *A*' was part of an aleatory rigging in Ω, we roll back in time, so to speak, to a point before the events occurred, and assume that they were generated by an aleatory rigging.

We can never know the exact magnitude of Π(*A*), unless we first know all facts of type *A* in Ω. But, to repeat, we can roughly infer the magnitude of Π(*A*). Of course, the precision of our approximation depends on the amount of facts we know in Ω. The more facts we know, the better approximation our estimate of Π(*A*) will be.

Now, assume instead that we are dealing with events of type *A*, in an event space Ω, partially constituted by historic time and historic facts, partly by future and events which have not yet occurred (and which must be unknown by definition). The principle for making inferences still remains the same. We use the historic facts we know in Ω to draw conclusions about the magnitude of Π(*A*). This, of course, also presupposes that *A* can be assumed to be generated by an aleatory rigging, both in the historical and the future part of Ω.

Thus, the procedure for historical, but unknown, facts is in principle the same as for future events. First, we delimit an event space Ω, and define the event type *A*. We count the number of *A*-premises in Ω (the size of the population). We study whether or not the properties of the surroundings are stable, which is a presupposition for aleatory equiprobability – a necessary condition for an aleatory rigging. If such is the case, we proceed to count how many cases of *A* and *A*' we find in Ω – we collect known facts that is.

We use the facts at hand for making inferences to the *A*-premises in Ω, the contents of which are unknown, either because they lie in the future and therefore lack content, or for the reason that the historical content has not been registered and preserved.

This procedure leads to a more or less precise estimation of the aleatory probability of *A* in Ω. The more *A*-premises in Ω we know the contents of, the better our estimation of Π(*A*) will be. In other words: The more facts of the type *A* (and *A*') we have in Ω, the better we can estimate the aleatory probability of *A* in Ω.

10. Evidence

The set of facts of type A (and A') which we know in Ω, and upon which we found our inferences, we shall call the *evidence set* $e = \{e_1, e_2, \dots, e_n\}$, where the elements $e_1, e_2, \dots, e_n$ constitute the *evidence*. Let us define the *evidence value operator* $\varepsilon(e_i)$. For ε , it holds that

(10.1.) $\varepsilon(e_i) = 1$ if e_i is a known fact A ;
$\varepsilon(e_i) = 0$ if e_i is a known fact A'.

Thus, the evidence value operator is the ”factual” correspondent to the Bernoulli variable X, which we defined for the outcomes of aleatory riggings. The evidence values may be arranged in an evidence vector

(10.2.) $\mathbf{E} = [\varepsilon(e_1), \varepsilon(e_2), \dots, \varepsilon(e_n)]$.

We define the evidence function

(10.3.) $E(\mathbf{E}) = \sum_i \varepsilon(e_i)/n$; $i = 1, 2, \dots, n$.

where E works like an ordinary arithmetical mean (expectation) operator.[17] Thus, E expresses the proportion of events of type A of all the facts of type A and A' that we know in Ω.

Thus, inferences regarding $\Pi(A)$ are made from the evidence we have in Ω. Since the ordering of the elements in the evidence set e does not matter for our inferences (all events of type A are both equal in properties and equiprobable in an aleatory rigging), E contains as much relevant information for our inferential purposes as does e. Therefore it does not matter whether we use E or e, and since E is easier to deal with mathematically, it is also clearly to prefer.

[17] To distinguish the operator of the evidence function from an arithmetical mean in general, we shall use an italic E to denote the former, and a non-italic E for the latter.

11. Hypotheses

Possessing evidence is however not enough. We must also formulate *hypotheses* to which we can confront our evidence. It is necessary to carefully distinguish between two kinds of hypotheses.

The one category, which we shall call *generator hypotheses*, are propositions of the kind ”the aleatory probability for the event *a* is Π”. All aleatorily probable events can be said to have been generated in an aleatory probability process, with the aleatory probability, or the *generator probability* Π. If Π is unknown, we can always formulate hypotheses as to its magnitude. But if the events cannot be modelled with an aleatory rigging, we will not get much farther than to the mere formulation of the hypothesis.

The aleatory rigging guarantees repetition of equal events with equiprobability under *ceteris paribus* conditions. When such a rigging is at hand, we may proceed from formulating hypotheses to evaluating how reasonableness of the hypotheses. When subject to an aleatory rigging, generator hypotheses will be propositions of the kind ”the aleatory probability for events of type *A* in Ω is Π”.

Since the real generator probability Π is a number between zero and unity, our guesses as to the value of the generator hypotheses concerning Π must also dwell in the range from zero to unity. In that way, generator hypotheses may be regarded as variables, who take on values between zero and unity in R^1. We shall denote this kind of variable by *G*. Thus, it is true that $0 \leq G \leq 1$.

A generator hypothesis is either true or false. It is true if $G = \Pi$, and false if $G \neq \Pi$. Generator hypotheses are never aleatorily probable. The reason for that is that they are propositions, not events. A generator hypothesis might well be conceived as subjectively probable, if somebody should believe that *G* to a particular degree between zero and unity. Analogously, it is conceivable that a generator hypothesis would be logically probable, provided it was combined with some other proposition *p*, so that *p* implies that *G* is rationally believable to a certain degree between zero and unity. But a generator hypothesis can never be aleatorily probable. So far the generator hypotheses.

The other category of hypotheses, which we shall call *empirical hypotheses*, are propositions of two different kinds: (1) ”*a* will be the case”, where *a* is a future event, or (2) ”*a* was the case”, where *a* was an unknown fact. The former type we shall call *empirical–future* hypotheses, the latter *empirical–historical*. We shall denote empirical hypotheses by *H*. *H* is *not* a variable, but a symbol for ”fixed” propositions.

Empirical–future hypotheses are, as distinguished from generator hypotheses, neither true nor false. The events which the propositions concern still have not occurred, and we cannot state with certainty that they will occur. We can only say that it is aleatorily probable that *a* will occur. But that is not what an empirical–future hypothesis does. It states categorically that ”*a will* be the case”. Such a proposition will, in due time, become true or false – the former if *a* occurs, the latter if *a* does not happen. But which is the status of such a proposition, before it can be factually determined whether *a* really is the case?

Empirical–historical hypotheses (type 2) are, like generator hypotheses, either true or false. Either *a* was the case, as the hypothesis stated, or it was not. But when unknown facts constitute part of an aleatory rigging, we may reason ”as if” we were dealing with future events. We then regard the facts we know as a sample drawn from the ”urn of history”, and the facts which we do not know as the unknown contents of the ”urn”.

We simply disregard that we are dealing with facts, and boldly reason as if these events did not yet occur, or were ”drawn from the urn of history”. The difference in our arguments about historical and future events consists of the following: We know that certain, imperturbable proportions prevail in the historic ”urn”, and that these proportions constitute the aleatory probability we seek. But in the future urn, the proportions are not necessarily fixed (as long as the population is finite), despite that we know that a particular aleatory probability exists. In practice, this makes very little difference.

Obviously, and empirical hypothesis is not aleatorily probable. Only events can be aleatorily probable, and an empirical hypothesis, be it future or historic, is not an event per se. But, analogously to generator hypotheses, we may conceive that empirical hypotheses are subjectively and/or logically probable.

Moreover, I will argue, empirical hypotheses are *epistemically probable*. This is a unique property to empirical hypotheses – generator hypotheses are *not* epistemically probable. At this point, I must ask the reader for some more patience with the explanation of the exact meaning of this.

From this section, we should bear in mind
(1) that different types of hypotheses exist – generator hypotheses G and empirical hypotheses H ;
(2) that hypotheses never are aleatorily probable ;
(3) that empirical hypotheses, but not generator hypotheses, are *epistemically probable*.

To bring our arguments further in the direction towards the concepts of epistemic probability and epistemic weight, it is first necessary to define conditional Kolmogorov weights.

12. Conditional Kolmogorov weights; Bayes' theorem

Recall that aleatory probabilities are special cases of Kolmogorov weights, namely where a particular interpretation (”the propensity to occur”) has been added to the purely mathematical properties. An aleatory probability always is a Kolmogorov weight, but the reverse does not necessarily hold true. Not to repeat the procedure of definition, we will carry out the definition in terms of Kolmogorov weights. Exactly the same definitions may be applied analogously for aleatory (or any other quantitative kind of) probabilities.

Assume two subsets A and B of some suitable set Ω (A and B may, but do not have to, denote events). The *conditional Kolmogorov weight for A, given B*, is given by

$$(12.1.)\qquad Q(A \mid B) \;=\; Q(A \cap B)/Q(A)$$

with the continuous case correspondent

$$(12.2.)\qquad q(A \mid B) \;=\; q(A \cap B)/q(A)$$

If B_1, B_2, ... , B_n are mutually exclusive, possess positive Kolmogorov weights, and together exhaust the entire Ω, then it holds true for every A that

$$(12.3.)\qquad Q(A) \;=\; \Sigma_i\, Q(B_i) \cdot Q(A \mid B_i)$$

with the continuous case correspondent

$$(12.4.)\qquad q(A) \;=\; \int q(B) \cdot q(A \mid B)\, \mathrm{d}B$$

and, under those same conditions, Bayes' theorem holds

$$(12.5.)\qquad Q(B_i \mid A) \;=\; \frac{Q(B_i) \cdot Q(A \mid B_i)}{\Sigma_i\, Q(B_i) \cdot Q(A \mid B_i)}$$

with the continuous case correspondent

$$(12.6.)\qquad q(B \mid A) \;=\; \frac{q(B) \cdot q(A \mid B)}{\int q(B) \cdot q(A \mid B)\, \mathrm{d}B}$$

Analogously to the denotations in Bayesian theory, we introduce the following terms: The unconditional Kolmogorov weight $Q(B_i)$ we shall call the *prior weight*. In the continuous case, we shall call the unconditional Kolmogorov density $q(B)$ the *prior density*. The conditional Kolmogorov weight $Q(B_i \mid A)$ we shall call the *posterior weight*. In the continuous case, we shall call the conditional Kolmogorov density $q(B \mid A)$ the *posterior density*.

As long as we take into consideration the conceptual relation between Kolmogorov weights and aleatory probabilities, we may boldly mix them in the same expression. The components in Bayes' theorem, for example, may partly consist of Kolmogorov weights, and partly of aleatory probabilities.

13. The likelihood function

Let us assume that we have access to certain evidence E from an aleatory rigging. If we formulate a generator hypothesis, i.e. that we choose a value of G, then we can always compute an aleatory probability for the evidence at hand having been generated in a process with the probability G. The process is a binomial process if the population of the rigging is ”large”, and a hypergeometric process if the population is ”small”.

The procedure is to compute a conditional aleatory probability for E, given G, which may be written $\Pi(E \mid G)$. The value of this aleatory probability always depends on three factors: (1) The value of the generator hypothesis G, (2) the number of ”trials” n (= the number of elements in the evidence set e), and (3) the relative frequency of ”favourable” outcomes E (= the number of events type A in e divided by n). In case we have a ”small” population, (4) the size of the population m (= the number of A-premises in Ω) must also be taken into consideration.

Thus, the conditional probability $\Pi(E \mid G)$ is a function of three (and four, respectively) variables. This function we shall call the *likelihood function*, and we shall denote it by $\Lambda(\ )$.

We thus distinguish between two cases: (1) *The discrete case*, where the population is ”small”, and where the likelihood function obeys a hypergeometric distribution in G, E, m and n.

$$(13.1.) \qquad \Lambda(G, E, m, n) \ = \ Q(E \mid G, m, n) \ = \ \frac{\left[\begin{matrix} G_i \\ nE \end{matrix}\right]\left[\begin{matrix} m - G_i \\ n - n \cdot E \end{matrix}\right]}{\left[\begin{matrix} m \\ n \end{matrix}\right]}$$

where G_i denotes the i:th of the (m+1) possible generator hypotheses. In the discrete case, namely, G is a discrete variable, which may assume (m+1) different values: $0/m$, $1/m$, ... , $((m–1)/m)$, m/m.

Case (2) is *the continuous case*, where the population is ”large”, and the likelihood function obeys a binomial distribution in G, E and n.

$$(13.2.) \qquad \Lambda(G, E, n) \ = \ Q(E \mid G, n) \ = \ \left[\begin{matrix} n \\ nE \end{matrix}\right] \cdot G^{nE} \cdot (1 - G)^{n-nE}$$

where G is a continuous variable, $0 \leq G \leq 1$.

14. The maximum likelihood-method, unbiasedness

We will now introduce a distinction between three different kinds of generator hypotheses. When G is supposed to assume a unique value, we shall speak about a *point hypothesis*. When the value of G is supposed to lie within an interval (between a lower bound $G_0 \geq 0$ and an upper bound $G_1 \leq 1$, where $G_1 > G_0$, we shall speak of an *interval hypothesis*. When several hypotheses are conjoined by disjunctions (logical ”or”–operators $\vee$), we shall speak of a *composite hypothesis*.

A common procedure to distinguish the ”best” of all conceivable point hypotheses is to seek out that hypothesis which yields the largest possible probability for the sample mean E which we did obtain, i.e. to seek the point hypothesis which has the largest likelihood value of all. The is R.A Fisher’s *Maximum Likelihood method* [the ML method].[18]

This point hypothesis may easily be obtained by the usual methods to find the global maximum of functions. The ML hypothesis G_{ML} is the value of G at which the likelihood function reaches its maximum value. That is

$$(14.1.) \qquad G_{ML} \;=\; \{G : \; \partial\Lambda(G, E, n)/\partial G \;=\; 0 \;\wedge\; \partial^2\Lambda(G, E, n)/\partial G^2 \;<\; 0\}$$

in the continuous (derivable) case. In the discrete case the same argument applies in principle, only another algorithm must be applied to find the maximum. It can be shown that the ML hypothesis G_{ML} always coincides with the sample mean obtained E. That is

$$(14.2.) \qquad G_{ML} \;=\; E$$

The ML method is generally considered to have the advantage to often yield ”unbiased” estimates. Unbiasedness means that if we have a large number of ”ML-estimators” G_{ML} from the same population, then the expected value of these estimators, $\mathrm{E}(G_{ML})$ will coincide with the true Π of the population.

But the principle of unbiasedness is a doubtful story. For assume that we have k ML estimates $G_{ML,\,j}$ from the same population, and that each of these are a function of a sample $[E_j, n_j]$ (in the continuous case) or $[E_j, m_j, n_j]$ (in the discrete case), where $j = 1, \ldots , k$. Then we can always ”pool” the samples by summing to $[\sum E, \sum n]$ (in the continuous case) or to $[\sum E, \sum m, \sum n]$ (in the discrete case), where $\sum E \;=\; \sum_j n_j{\cdot}E_j/\sum_j n_j$; $\sum m \;=\; \sum_j m_j$; $\sum n \;=\; \sum_j n_j$.

[18] The method was allegedly used by both Daniel Bernoulli and Karl Friedrich Gauss (according to Hacking 1965, 176), but it is named after and associated to Fisher. The standard reference Fisher (1925).

In other words, the very notion of counting expected values from several ”estimates” from the same population is self-contradictory. Our quantity of information must always be the *total* information we get from *all* such ”estimates”. The procedure to chop this quantity of information up and compute ”average estimates” by using the parts contains a self-contradiction. We cannot both have access to all these ”estimates” and be unable to ”pool” the to a common information background, from which we obtain a new estimate.

Therefore, the principle of unbiasedness is not applicable, and unbiasedness cannot constitute a ”guiding principle”, according to which we judge whether a method of estimation is good or poor.

Let me give a simple example. We have the customary urn with a number of black and white balls in unknown proportions. Attach unity value to ”black ball” and zero value to ”white ball”. Now assume that we have drawn three balls from the urn, so that $E = 1$ and $n = 3$. Then the ML estimate, which is ”unbiased”, would be $G_{ML} = 1$.

For purely intuitive reasons, I am sceptical to the proposition that the ”best” hypothesis would be that $G = 1$ in case we have drawn three balls from the urn. Personally, I would not at all suggest that $G = 1$ until I had drawn a rather large number of balls and found them all black. But this raises the question whether this intuitive aversion to the ML method’s suggestion of ”best” hypothesis can be given a more solid foundation in rational arguments?

A characteristic of the ML method is that the choice of ”best” hypothesis mirrors a particular aspect of the curvature of the likelihood function. Where the vertex of the likelihood function is located, there dwells the ”best” hypothesis, according to this line of thought. But the question is why this characteristic should be decisive?

If we want our ”best” hypothesis to reflect the position of the ”likelihood mass” (the area beneath the likelihood curve), in my opinion it would be more natural to choose the *average* likelihood value than the maximum. In that case, an *Average Likelihood method* [AL method] would be more reasonable than the ML method, to characterise the hypothesis which ”best” reflects the likelihood mass.

But not even the AL hypothesis needs be the ”best”. We may have particular reasons to believe that some hypotheses in the interval [1, 1] are more reasonable than others, regardless of which evidence we obtained from our trials. Therefore, we may conceive giving some hypotheses a larger weight than others in our estimate, so that we use a *Weighted Average Likelihood method* [WAL method], in which the array of *hypothesis weights* obey the

Kolmogorovian axioms, and hence are Kolmogorov weights. The AL method will then appear as a special case of the WAL method, namely when all hypothesis weights are equal.

Thus we may conceive several different methods built on the notion that the ”best” hypothesis is obtained by studying the properties of the likelihood function. On this background it may be objected that it is not the conditional probability of the likelihood function $\Pi(E \mid G)$ which is our ultimate interest to obtain, but the ”inversely conditioned” probability $\Pi(G \mid E)$.

For what we are seeking is not the hypothesis which renders our evidence (the sample we happened to obtain) the most probable.[19] Instead we are seeking the hypothesis which is the most probable, given our obtained evidence.

[19] ”Most probable” should be understood in a loose sense and not be interpreted in terms of maximisation.

15. The Bayesian approach – two problems

To obtain the ”inverse” probability (or posterior probability) we are seeking, we must ”reverse the conditioning”, which can only be done by applying Bayes’ theorem. In the discrete case, we get

$$(15.1.)\qquad \Pi(G_i \,|\, E) \;=\; \frac{\Pi(G_i) \cdot \Pi(E \,|\, G_i)}{\sum_i \Pi(G_i) \cdot \Pi(E \,|\, G_i)} \;=\; \frac{\Pi(G_i) \cdot \Lambda(G, E, m, n)}{\sum_i \Pi(G_i) \cdot \Lambda(G, E, m, n)}$$

since $\Pi(E \,|\, G_i) \;=\; \Lambda(G, E, m, n)$.

Here already we discern a serious problem, namely $\Pi(G_i)$. We did already establish that anything like $\Pi(G_i)$ does not exist – hypotheses are propositions, not events. Only events can be aleatorily probable. Therefore, hypotheses cannot be aleatorily probable, and that is exactly why there is no such thing as $\Pi(G_i)$.

But if $\Pi(G_i)$ does not exist, how can we ”reverse the conditioning”? The only feasible way is to use the possibilities to mix aleatory probabilities with other kinds of Kolmogorov weights in Bayes’ theorem. When we to reverse the conditioning, using that procedure, we obtain

$$(15.2.)\qquad Q(G_i \,|\, E) \;=\; \frac{Q(G_i) \cdot \Pi(E \,|\, G_i)}{\sum_i Q(G_i) \cdot \Pi(E \,|\, G_i)} \;=\; \frac{Q(G_i) \cdot \Lambda(G, E, m, n)}{\sum_i Q(G_i) \cdot \Lambda(G, E, m, n)}$$

The result of the procedure – the inversely conditioned posterior weight $Q(G_i \,|\, E)$ – is merely a Kolmogorov weight, not an aleatory probability. For when we mix aleatory probabilities with Kolmogorov weights in Bayes’ theorem (or other arithmetical expressions where this is feasible), ”least common denominator” will be determining the outcome property.

All aleatory probabilities are also Kolmogorov weights, but not the reverse. Thus we cannot take it that an arithmetical operation results in an aleatory probability *unless all other magnitudes in the operation are also aleatory probabilities*.

Since our input values (the right-hand side of the Bayes’ theorem equation) are partly aleatory probabilities, partly Kolmogorov weights, the result of the operation (the left hand side of the equation) cannot express an aleatory probability. The ”least common property” – that all input values are Kolmogorov weights – will determine property of the output.

The fact that a the outcome of reversing the conditioning is a Kolmogorov weight, and not (necessarily) a probability, brings on some interpretation problems of philosophical nature. These interpretation problems are common for both the discrete and the continuous case. But in the latter case there is another problem, which we first we first must describe and solve. Therefore we shall leave the interpretation problem for the time being and get back to it later.

The continuous case version of Bayes' theorem is

$$(15.3.) \qquad q(G \mid E) \;=\; \frac{q(G) \cdot \pi(E \mid G)}{\int q(G) \cdot \pi(E \mid G)\, \mathrm{d}G} \;=\; \mathbf{?}$$

The question mark indicates another problem, which only appears in the continuous case. In the discrete case, we substituted the likelihood function both in the numerator and the denominator of the right-hand side of the Bayes' theorem equation. This cannot be done in the continuous case, since the likelihood function $\Lambda(G, E, n) \;=\; \Pi(E \mid G)$ expresses an aleatory *probability*, not an aleatory *density*. But Bayes' theorem includes an aleatory *density*, not and aleatory *probability*. How can we get by this *dimensional problem*?

The aleatory probability for our sample, as expressed in the likelihood function, depends on which point hypothesis we formulate. As we have seen, the likelihood value expresses an ”ordinary” aleatory probability, not a density. This causes concern in the continuous case, when we have an infinite number of conceivable hypotheses in the interval [0, 1].

For this reason a point hypothesis cannot be given an ”ordinary” prior weight $Q(G)$, but only a prior density $q(G)$.

Alternatively, we may formulate an interval hypothesis, and give it an ”ordinary” prior probability, but this procedure fails since it will not render a finite likelihood value corresponding to that hypothesis. Why is that, then?

The likelihood function is a continuous function, the definitional set of which ranges from zero to unity, namely the various feasible point hypotheses G about the Π probability of the population. The value set contains (conditional) probabilities. It is important to grasp that these are perfectly ordinary probabilities with a ”probability mass” of their own – they are not probability densities. It is tempting to prematurely conclude that the likelihood function, being a continuous function, expresses probability densities. But such is not the case.[20]

The implication of this is that we *cannot* form ”likelihood intervals” corresponding to interval hypotheses, because each interval hypothesis will have an *infinite* likelihood value. This may best be illustrated by computing the likelihood value for a composite hypothesis, the components of which lie within the interval we are interested in.

For example, let us assume that we are interested in the interval hypothesis ”$0,5 \leq G \leq 0,7$”. Now, form the composite hypothesis ”$G = 0,5 \vee G = 0,7$” The likelihood value for this hypothesis is

(15.4.) $\Lambda(0,5\ ; E, n) + \Lambda(0,7\ ; E, n)$

Let us now expand the composite hypothesis by adding more points into it, for example ”$G = 0,5 \vee G = 0,55 \vee G = 0,6 \vee G = 0,65 \vee G = 0,7$”. The likelihood value of this hypothesis is, correspondingly,

(15.5.) $\Lambda(0,5; E, n) + \Lambda(0,55; E, n) + \Lambda(0,6; E, n) + \Lambda(0,65; E, n) + \Lambda(0,7; E, n)$

Now we see that the likelihood value will increase rapidly the more we ”comb” the interval with conjunctions of point hypotheses. This shows that we cannot integrate the likelihood function to compute ”likelihood intervals” corresponding to interval hypotheses.

The mere fact that the likelihood value swiftly exceeds unity when the ”teeth” of the hypothesis ”comb” are condensed demonstrates that it is neither very meaningful to reason in terms of interval hypotheses not in terms of composite hypotheses. Really, it is only the simple point hypotheses who seem reasonable. The composite hypotheses only fill the function to illustrate this argument, and the interval hypotheses aid – as we are about to see – in the redefinition of the concept of ”point hypothesis”, so that it gets a meaningful interpretation in the continuous case as well.

Thus there is a danger in mixing ”ordinary” *Q*-weights (or probabilities) with *q*-densities (or probability densities) without reflection. To straighten this out, it is necessary to reformulate the notion of point hypotheses. Instead of letting a point hypothesis denote one single

[20] *Vide* e.g. Hacking (1965), chapter XI, for a discussion of this. Hacking claims that ”the likelihood value does not obey the Kolmogorovian axioms”, which is an erroneous proposition, founded on the confusion caused by failure to distinguish densities from probability masses. The likelihood value is a conditional probability, and conditional probabilities always obey the Kolmogorovian axioms.

numerical value G, we must regard the point hypothesis as a limiting value to an interval hypothesis $G_0 \leq G \leq G_1$, when the width of the interval $\Delta G = G_1 - G_0$ approaches zero. Such a limiting value is preferably defined for the lower bound of the interval G_0, which gives us the following expression for the likelihood value of a point hypothesis $G = G_0$:

$$(15.6.) \qquad \lim_{G_1 \to G_0} \Lambda(G_0 \leq G \leq G_1, E, n) \;=\; \lim_{G_1 \to G_0} \Pi(E \,|\, G_0 \leq G \leq G_1, n) \;=$$

$$= \lim_{\Delta G \to 0} \Pi(E \,|\, G = G_0 + \Delta G, n) \;=\; \pi(E \,|\, G_0) \;=\; \lambda(G_0, E, n)$$

where $\lambda(G_0, \bullet\,)$ expresses the *likelihood-density* at G_0. By using point hypotheses in their capacity as limiting values to interval hypotheses, and likelihood-densities, the probability density for the point hypothesis will turn out ”dimensionally compatible” with the likelihood value. Let us apply this on Bayes’ theorem in the continuous case. We get that

$$(15.7.) \qquad q(G \,|\, E) \;=\; \frac{q(G) \cdot \pi(E \,|\, G)}{\int q(G) \cdot \pi(E \,|\, G)\, \mathrm{d}G} \;=\; \frac{\pi(G) \cdot \lambda(G, E, n)}{\int \pi(G) \cdot \lambda(G, E, n)\, \mathrm{d}G}$$

The dimensional problem in the continuous case can thus be considered as solved, by going from using a likelihood mass Λ to using likelihood densities λ in the expressions. But the former problem still remains – to give the prior weight (the prior density) and the posterior weight (the posterior density), respectively, an interpretation in terms of probability. I must ask the reader for further patience with this question.

16. The WAL method is Bayesian!

Of course, we are not only interested in one single posterior weight $Q(G_0 \mid E)$ (or $q(G_0 \mid E)$, in the continuous case [21]), corresponding to one single point hypothesis. We are seeking the whole spectrum of posterior densities $Q(G \mid E)$ for all conceivable values of G, ranging from zero to unity.

This ”spectrum” of posterior weights we shall call the *posterior distribution*. The shape of the posterior distribution depends on two categories of factors. One is the array of evidence [*E*, *m*, *n*]. The other is our choice of prior weights, or *hypothesis weights*, for the different hypotheses. Our choice of hypothesis weights defines a ”spectrum” of prior weights, which we shall call the *prior distribution*.

The shape of the posterior distribution thus partly depends on which prior distribution we use, partly on which array of evidence we confront this prior distribution. The function of the prior distribution is simply to weigh the evidence, as expressed by the likelihood value, by different hypothesis weights. The result of this weighting is a certain posterior distribution.

The posterior distribution consists of Kolmogorov weights. Hence it has, like other distributions of Kolmogorov weights (i.e. what is usually called ”probability distributions in statistics textbooks), properties like location and dispersion.

When we compute the hypothesis-weighted mean over all conceivable values of the generator hypotheses, we obtain a unique value, and that unique value is nothing but what we have previously called a WAL estimate of Π.[22]

[21] Not to encumber the presentation, I will only show the discrete case in the text. I leave it to the reader to draw parallels to the continuous case.

[22] This, of course, presupposes that the sum (the integral) is finite, so that such a weighted average exists.

Let $\alpha(G)$ denote the frequency function of the prior distribution (the array of hypothesis weights), and $\zeta(G)$ the frequency function of the posterior distribution.

In the discrete case, it holds true that

$$(16.1.)\qquad G_{WAL} = \Sigma_i\, G_i \cdot \zeta(G_i)\,,$$

where

$$(16.2.)\qquad \zeta(G_i) = Q(G_i \mid E) = \frac{Q(G_i) \cdot \Pi(E \mid G_i)}{\Sigma_i\, Q(G_i) \cdot \Pi(E \mid G_i)} = \frac{Q(G_i) \cdot \Lambda(G, E, m, n)}{\Sigma_i\, Q(G_i) \cdot \Lambda(G, E, m, n)}$$

In the continuous case, it is true that

$$(16.3.)\qquad G_{WAL} = \int G \cdot \zeta(G)$$

where

$$(16.4.)\qquad \zeta(G) = q(G \mid E) = \frac{q(G) \cdot \pi(E \mid G)}{\int q(G) \cdot \pi(E \mid G)\,\mathrm{d}G} = \frac{\pi(G) \cdot \lambda(G, E, n)}{\int \pi(G) \cdot \lambda(G, E, n)\,\mathrm{d}G}$$

It is perfectly correct to say that the WAL estimate is based on a Bayesian procedure. The WAL estimation involves that we (1) compute a Bayesian posterior distribution, based on the distribution of hypothesis weights (the prior distribution) which we find suitable, and (2) compute the arithmetical mean of the posterior distribution obtained.

Thus the WAL estimate is nothing else than the mathematical expectation of the posterior distribution, i.e.

$$(16.5.)\qquad G_{WAL} = \mathrm{E}[\zeta(G)]$$

The same array of evidence [E, m, n] may certainly give rise to different WAL estimates, depending on the scheme of weights applied for the hypothesis weighting, i.e. which prior distribution we apply. Therefore there is reason to name the WAL estimates after the prior distribution chosen. If the prior distribution is $\alpha(G)$, we shall say that the WAL estimate is *α-weighted*.

As mentioned above, the AL method is a special case of the WAL method, namely where equal hypothesis weights are applied, i.e. that $\alpha(G) \in Re(0, 1)$. We then say that the WAL estimate is *unweighted*, since an unweighted arithmetical mean is being applied.

17. Dispersion of posterior distribution, evidence weight

As previously noted, the posterior distribution, like other distributions, has properties like location and dispersion. A suitable way to describe these properties is to compute indicators of location and dispersion.

The WAL method achieves one of the two – it gives us the expected value of the posterior distribution, which is an excellent location indicator. But the WAL method does not render us any idea of the dispersion of the posterior distribution.

The most common measure of dispersion is the variance (and the dimensionally adjusted standard deviation). To complete the description of the properties of the posterior distribution, it may be suitable to compute its variance, the *posterior variance,* which is defined by

(17.1.) $\mathrm{V}[\zeta(G)] = \mathrm{E}[\zeta^2(G)] - \mathrm{E}^2[\zeta(G)]$

and the *posterior standard deviation* $\mathrm{D}[\zeta(G)] = \sqrt{}\mathrm{V}[\zeta(G)]$.[23]

When the same prior distribution is being used consistently in weighting evidence from a certain aleatory rigging, the variance (and hence also the standard deviation) will always be diminishing when the number of elements in the evidence set increases. In other words, an increase in the quantity of information will reduce the dispersion of the posterior distribution.

A case of little practical importance, but which is very important in principle, is when the evidence set is empty, i.e. when $e = \varnothing$. Even in this case the posterior distribution will often (I say ”often” because the choice of prior distribution is crucial) have a definite expected value and a finite variance. This is so, despite E not being computable (there are no terms by which to compute the expected value $E(\mathbf{E})$)! This strange phenomenon depends exclusively on the definitional peculiarity 0! = 1, which implies that

(17.2.) $\begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0!/[0!(0-0)!] = 1/1 = 1$

This expression occurs both in the discrete and the continuous version of the likelihood function, which guarantees that a likelihood value exists even when $e = \varnothing$. For the likelihood function, the following simply holds true

(17.3.) $\Lambda(G, e = \varnothing, m, n = 0) \in Re(0, 1)$

[23] The definitions and the ensuing argument presuppose that a prior distribution which allows these computations has been chosen. This proviso will cause no trouble in practice.

and corresponding λ in the continuous case. Under these conditions, the likelihood function lets the prior distribution ”right through”, so that the posterior distribution is congruent to the prior distribution. My expression ”often” may thereby be specified more precisely. The posterior distribution possesses an expected value and a finite variance even when $e = \varnothing$, if and only if the prior distribution possesses an expected value and a finite variance.

Which, then, is the importance in principle of this case? That question has two aspects. One aspect is philosophical – how do we interpret the posterior distribution and its properties when we have no evidence at all? Regarding this question, I must keep the reader curious for another while. The other aspect, which is purely mathematical, we can deal with right away.

The choice of prior distribution will affect the posterior variance. As long as we stick to the same prior distribution, it holds true that the posterior variance will shrink when the sample size grows (larger evidence set). When the evidence set is empty, the posterior variance assumes its maximum value. When the evidence set accommodates the entire population, i.e. $n = m$, (or $n \to \infty$ in the continuous case), so that we reach full knowledge about the whole population, then the posterior variance will vanish. That is

(17.4.) $\mathrm{V}[\zeta] = 0 \quad \text{if } n = m$

in the discrete case, and

(17.5.) $\lim_{n \to \infty} \mathrm{V}[\zeta] = 0$

in the continuous case, respectively.

This is true regardless which prior distribution we choose (within reasonable limits, set by properties like convergence of integrals). All imply that the posterior variance converges towards zero when our evidence ”exhaust” the population. All prior distributions with finite variance $\mathrm{V}[\alpha]$ yield this very posterior variance when the evidence set is empty – i.e. $\mathrm{V}[\zeta] = \mathrm{V}[\alpha]$, provided that $e = \varnothing$.

The posterior variance is near at hand to use as an ”information measure”,[24] which reflects the balance between what we know, and what we do not know. There are however two problems in this context: (1) The variance is not a standardised measure – the maximum variance

[24] Many other information measures may be construed. In this connection we will concentrate on functions of the posterior variance.

(provided that it exists) always lies considerably below unity, and (2) the variance goes ”the wrong way” – it goes down when *n* increases. An information measure ought to increase when the information increases, not the other way around.

Thus the desirable properties of the magnitude we seek are the following: The magnitude should (1) be a monotonously decreasing function of the posterior variance, (2) assume zero value when the evidence set is empty, (3) assume unity value when the evidence set accommodates the entire population (converge to zero when $n \to \infty$ in the continuous case). Moreover it is an advantage is the measure is dimensionally adjusted, so that it mirrors the standard deviation rather than the variance. A square root function of the variance thus seems suitable.

The chief problem of this approach is that we must find a method to standardise the prior variances. Since they are different depending on which prior we use, they yield different values of the posterior variance when the evidence set is empty.

But as long as the prior variance is finite, we can use (the inverted value of) this *maximum posterior variance* $\mathrm{V}^{-1}[\alpha]$ as an adjustment factor when computing the measure we are looking for. By this procedure, the mathematical product of the adjustment factor and the prior variance will always be unity, because $\mathrm{V}^{-1}[\alpha] \cdot \mathrm{V}[\alpha] = 1$.

Since the adjustment factor is a constant, it also holds true that the mathematical product of the adjustment factor and the posterior variance converges to unity as the evidence set exhausts the population. Thus the adjusted posterior variance always goes from unity (when the evidence set is empty) to zero (when the evidence set exhausts the population). Hence, the adjusted posterior standard deviation (whose adjustment factor is $\mathrm{D}^{-1}[\alpha] = \surd\mathrm{V}^{-1}[\alpha]$) also has these properties.

The measure we are seeking, and which we shall call the *weight of the evidence* with respect to the WAL estimate (denoted by W_{WAL}, may easiest be defined as unity minus the adjusted posterior standard deviation, that is

$$(17.6.) \qquad W_{WAL} = 1 - \mathrm{D}^{-1}[\alpha] \cdot \mathrm{D}[\zeta]$$

The measure characterises our information position and thus the degree of precision[25] of our WAL estimate, given the chosen distribution of hypothesis weights (the prior distribution

[25] ”Precision” should be understood in a loose sense. We are not talking about the measure which is usually called precision, and which is defined as the inverted value of the variance V^{-1}.

$\alpha(G))$. That is, W_{WAL} measures the bearing of the evidence on the dispersion of the posterior distribution *in relation to the dispersion of the used prior distribution*. Strictly speaking, W_{WAL} does not measure our information position as a whole, but merely the *additional information* provided by our evidence.

It was Keynes who first suggested that the quantity of evidence (”the argument”) supporting a logical probability should be called ”the weight of the argument”.[26] This is an important part of Keynes’s theory of logical probability. Even if our theory still moves on an abstract, purely mathematical level, where we confine the interpretations of the Kolmogorov weights to regard the likelihood value as an aleatory probability, there are parallels between the measure we defined – the weight of evidence – and what Keynes calls ”the weight of an argument”, clear enough to make suitable the naming of our concept after that of Keynes’s. [27]

We must make clear to ourselves the difference between the concentration around the expected value (the WAL estimate) of the posterior distribution on one hand, and the weight of the evidence. The concentration of the posterior distribution reflects two things: (1) the choice of prior distribution and the concentration of that chosen distribution, and (2) our evidence and the ”*contribution* to concentration” which they bring about. The weight of the evidence, on the contrary, only refer to (2), not to (1).

Both the location of the posterior distribution (the WAL estimate as such) and its dispersion are affected by the choice of prior distribution. Thus that choice is significant to the estimates we obtain, and for that reason it must be subjected to closer scrutiny.

[26] *Vide* Keynes (1921), chapter 6.
[27] Excellent presentations of Keynes’s argumentation can be found in Runde (1990, 1991).

18. The choice of prior distribution

The likelihood function can be said to ”encapsulate” the information, or the knowledge, that we obtain by drawing a sample and studying the obtained evidence. The likelihood function thus encapsulates an information *addition*, or the difference between the information position before and after the acquisition of evidence.

The information situation before the acquisition of evidence we shall call the *prior information* $\mathbf{I}_\alpha$, , and the information situation after the acquisition of evidence we shall consequently call the *posterior information* $\mathbf{I}_\zeta$. It holds true that $\mathbf{I}_\zeta = \{\mathbf{I}_\alpha \cup e\}$.

It is desirable that we choose our prior distribution in such a way that it reflects our prior information. In that way our posterior distribution will reflect our posterior information, too. The problem of choosing a prior distribution may thus be expressed as the quest for the *information function* ϕ , which transforms the prior information $\mathbf{I}_\alpha$ to a prior distribution $\alpha(G)$.

(18.1.) $\alpha(G) = \phi(\mathbf{I}_\alpha)$

An important philosophical question, which has to be cleared up before we can proceed to erect a system where prior and posterior distributions reflect the respective information positions, is whether there exists such a thing as a unique information function. The choice of prior distribution thus implies that we must take on the difficult issue of how to philosophically interpret the Kolmogorov weights (-densities) of which the prior distribution consists.

But there is also another, practical aspect of the choice of prior distribution. As is it the case that the larger the sample (the evidence set), the less important is the choice of prior distribution to the location and dispersion of the posterior distribution. De various distributions tend to converge when the sample grows large.

For that reason it is seldom necessary in practice to spend much mental effort on the choice of prior distribution. One usually picks a suitable distribution which roughly fits with the prior information, and which is mathematically convenient to handle. Such prior distributions are preferably chosen within *conjugated* families of distributions.[28] In our continuous case, where the likelihood function is binomially distributed, a prior distribution ought to be chosen from the family of beta distributions. [29]

[28] We will not enter into a thorough discussion of these mathematical properties. Definitions and a discussion of conjugated families of distributions are found in e.g. de Groot (1970), chapter 9.

[29] For a definition of the beta distribution, *vide* e.g. Hogg and Tanis (1983).

But the solution of the practical problems do not imply any solutions of the philosophical queries. Even if we should use a certain, conjugated prior distribution for the sake of convenience, the fact still remains that we try to approximate our prior information by using it. For such a procedure to be justified, we must first, to repeat, clear up whether we really can characterise prior information by a prior distribution. To *approximate* the information function both presupposes that it exists, and that we know its shape.

19. Facts, logic and the information function

When analysing the information function, we take particular interest in two problem complexes. (1) What is really meant by the ”prior information” (and the ”posterior information”)? Are these concepts subjective or objective? ; (2) How can we know that a certain array of prior information yields one and only one specific prior distribution? (the question of the existence of the information function), and – if it exists – How can we know the shape of the information function? Let us take these questions on, one by one.

By ”information” we mean *acquired factual knowledge*. Endowed knowledge (an infant knows how to breathe without getting any instructions) are not counted into this category. Nor do we count intuition – like an excellent mathematician can find a correct proof without having seen it first – or talent – like a musician with perfect pitch can tune up a G just like that. Factual knowledge are concerned with what is the case, and what is the case cannot be known until it really has been the case. Factual knowledge must be acquired, it is empirical knowledge. It is that kind of knowledge we are speaking of when using terms like ”information” or ”knowledge” in our context.

Facts are objective. The very word ”fact” refers to what is the case. It is not enough that one person, or even many persons, *regard* a phenomenon *F* to be the case for it to be established as a fact *F**. Something more is required, namely an *examination* according to some established ethical code *C*, for a phenomenon to be established as a fact. When a phenomenon has passed such a scrutiny, and thus is established as a fact, it does not matter how many persons who regard *F* a fact. It may well be that all are touchingly unanimous that *F* is not a fact.

If *F* has been established as a fact, and this is disputed by somebody (or by many), then it is not only disputed that *F**, but also the very code of scrutiny *C*. Provided *C* is right, and that *C* has been correctly applied to *F*, then *F** will stand fast regardless of how many who question it.

A drastic, but illuminating, example, is the French historian Robert Faurisson’s denial of the existence of gas chambers in Auschwitz. Let *F* denote the proposition that they did exist. Now, it is an established fact *F** that the gas chambers did exist. Let us say that this fact has been established by the code *C*. What Faurisson must show is that *C* is inadequate, or that *C* has been inadequately applied to *F*. Of course Faurisson is wrong – he is not able to show any of these. That Faurisson happens to have followers does not alter the case. Even if the whole human race would deny that *F*, it still remains that *F**. It is an absurd thought that the whole human race would dismiss *C*, and it is equally inconceivable that the thorough documentation of *F**, by *C*, would be generally rejected.

All accumulated facts cannot be possessed by individuals, an obvious statement considering the existence of libraries and databases. When correct facts are stored they form a pool of objective experience. Such a pool constitutes a kind of collective memory bank. Historical science, which is an important part of any scientific discipline, spends much of its efforts to gather facts for such collective experience banks. This is not done indiscriminately, the aim of the sifting is to only give real, true facts access to such banks. The information stored are taken for objective facts, those who are discarded must live in the shadowy world of the probable.

I will not go deeper into the intricate questions of ethics in science associated to this sifting. But generally it can be stated that the question what should be established as objective facts cannot properly be viewed as a matter of purely subjective considerations. Each scientific discipline has its own code in this respect, and there are always unspoken or tacit rules on how the discernment should be carried out. This codes have been laboriously established within each separate discipline, and it may be difficult to find any general patterns of ethical rule stretching across disciplinary boundaries.

Our prime interest in this context is however not the formulation of these rules. We are content with the existence of such codes, and that they are applied to discern and establish objective facts.

Factual knowledge can only refer to such objective facts. When we speak of knowledge (or ”information”), it is only allowed to refer to facts, not to any other personally or generally held beliefs.

The prior information is the set of information we refer to in order to motivate the choice of a particular prior distribution. This information mass may be large or small, as it pleases us. The important thing is that we state *which* information we are referring to. The concept of information is objective – it deals with facts in the collective experience bank. We need not take all facts into consideration, but only the facts that are *relevant* to the problem at hand, namely to estimate a particular aleatory probability $\Pi(A)$ of a certain aleatory rigging in the event space Ω.

The question of relevance is just as important in this context as it was to the rigging of the aleatory process. What do we really mean by saying that facts are *relevant*? A reasonable definition is that facts are relevant if they affect the prior distribution. But is this not begging the question? For how can we know what does, and what does not, affect the prior distribution?

When we say ”affect”, we refer to a *causal relation*. The information function must refer to a causal relation. We conceive that the prior information affects the prior distribution, and that this affectation may be described by the information function ϕ.

It is a difficult philosophical question to determine what a causal relation really is, and whether or not we can *know* that one phenomenon causes another. The British 18:th Century philosopher David Hume argued that we never really can know whether causal relations really exist.[30] This is the famous scepticism of Hume’s. If Hume was right, that seems to imply that we can never be certain about anything like an information function. Let us examine whether things really are that bad.

The intricate philosophical questions about the concepts of cause and effect are closely related to the theory of probability. However, going deeper into that complex of questions would burst the frames of this essay. We must make do with the conclusion that causes are generally dealt with using the same kind terms as with probabilities.

For example, we distinguish between *real* and *known* causes [*causa essendi* and *causa cognoscendi*, respectively. This distinction corresponds to that between aleatory and epistemic probabilities. The former refers to the ”propensity to occur” of an event, the latter to *what we know* about that ”propensity to occur”.

But causes may also be conceived to be subjective. Let us call these *probable causes*. The corresponding concept in probability theory is ”the degree of belief”, or subjective probability. We may also conceive *logical* causes as an objective concept, corresponding to the ”rational degree of belief” of the logical theory of probability.

When dealing with the information function it is important to distinguish between these different categories, and the arguments associated with each of them. $\phi(\mathbf{I}_\alpha) \rightarrow \alpha(G)$ can be given different interpretations, even if we assume that the prior information $\mathbf{I}_\alpha$ is the same set in all cases.

[30] The argument is to be found in Hume (1748).

The *subjective interpretation*[31] is that the consideration of $\mathbf{I}_\alpha$ implies that the individual or subject believes that $\alpha(G)$. This interpretation is problematic, not least because the same $\mathbf{I}_\alpha$ may lead to completely different $\alpha(G)$ depending on the subject. This, in a nutshell, is the reason why the subjective theory of probability has its limitations for scientific purposes. If I believe this, and you believe that – who is right? That question can only be answered if the query really is a matter of fact.

If the concept of probability refers to the ”degree of belief”, then in the end there are no matters of fact in the theory of probability. The probable is what we believe, and to reason scientifically about pure matters of belief may be very awkward. Experience has shown subjective probabilities to be most useful in axiomatic decision theory, where decisions are assumed to be governed by the agents’ ”degree of belief”, as well as their preferences facing different choice alternatives.

The *logical interpretation* is that $\mathbf{I}_\alpha$ implies $\alpha(G)$. Let us examine closer what this might mean.

Assume that our present aleatory rigging of $\Pi(A)$ in the event space $\mathbf{\Omega}_A$ is *equal* to another known (subset of $\mathbf{I}_\alpha$) aleatory rigging $\Pi(B)$ in the event space $\mathbf{\Omega}_B$. Then we know in advance that $\Pi(A) = \Pi(B)$. The result of this is of course that the investigation of $\Pi(A)$ is superfluous. We already know that $\Pi(A) = \Pi(B)$, and the only generator hypothesis we have is $G_A = \Pi(B)$. The prior distribution will then be a single-point distribution, where $Q(G_A) = 1$, and the posterior will consequently be that same. This case is obviously uninteresting.

Now assume that our present aleatory rigging of $\Pi(A)$ in the event space $\mathbf{\Omega}_A$ is a *subset* of another set of aleatory riggings $\Pi(\bullet)$ in the event spaces $\mathbf{\Omega}_\bullet$, where $\Pi(\bullet)$ lies in between two values Π_0 and Π_1 . It follows that $\Pi(A)$ also lies in between these two values. All generator hypotheses outside the interval $[\Pi_0, \Pi_1]$ are thereby excluded, and must be given zero weights. But how do we know that $\Pi(\bullet)$ lies in the interval $[\Pi_0, \Pi_1]$? We cannot, unless we *know* all riggings in $\Pi(\bullet)$ in $\mathbf{\Omega}_\bullet$, and if we know them, we also know $\Pi(A)$ in $\mathbf{\Omega}_A$ which is a subset of the former category.

More examples could be given. But the above should be enough to demonstrate the hopeless character of the task to logically deduce $\alpha(G)$ from $\mathbf{I}_\alpha$, at least by using ordinary two-valued logic. It is possible that there is a ”relation of partial implication” (RPI) from $\mathbf{I}_\alpha$ to $\alpha(G)$, but the question is which relation? And how large is ”the rational degree of belief” $\alpha(G)$ does $\mathbf{I}_\alpha$

[31] A fervent advocate of this interpretation is de Finetti. *Vide* e.g. de Finetti (1972, 1990).

bring about? No clear, practically useful theory has been constructed to deal with this problem, and I suspect that such a theory will never be constructed, for that matter. The reason for that might be that such a theory is plainly impossible to formulate.

So, even if there were a RPI going from $\mathbf{I}_\alpha$ to $\alpha(G)$, and that relation would be liable to the interpretation that the certainty of $\mathbf{I}_\alpha$ causes the rational degree of belief that $\alpha(G)$, the fact remains that we do not know this relation. If there really is a causal relation $\phi(\mathbf{I}_\alpha) \rightarrow \alpha(G)$, it is certainly not a *known* causal relation.

Let us now talk of *known* causes, about logical relations who *demonstrably* exist.

In our first example, with equal riggings, we do have that kind of relation. Equality is an ordinary, two-valued logical operator. If it is true that the rigging $\Pi(A)$ in $\Omega_A = \Pi(B)$ in Ω_B, where $\Omega_B \in \mathbf{I}_\alpha$, then it follows that all other values of G_A than $\Pi(B)$ must be incorrect. The only possible value of G_A is $\Pi(B)$. In this case, the ”rational degree of belief” that G_A is $\Pi(B)$ be unity, because it is a logical truth. Thus it is true that $\mathbf{I}_\alpha \Rightarrow \alpha(G_A)$, where $\alpha(G_A) = 1$ if $G_A = \Pi(B)$, and $\alpha(G_A) = 0$ if $G_A \neq \Pi(B)$. A logical information function ϕ can thus be defined in this case. But we have no practical use of it whatsoever, as the case is trivial.

In our second example it is not possible (at least not with the means that presently are at our disposal) to show that a logical implication $\mathbf{I}_\alpha \Rightarrow \alpha(G)$ exists. If we choose a pair of arbitrary values of G, say G_0 and G_1, we can neither show that $\mathbf{I}_\alpha \Rightarrow \alpha(G_0)$ nor that $\mathbf{I}_\alpha \Rightarrow \alpha(G_1)$, where $\alpha(G_0)$ and $\alpha(G_1)$ denote the values between zero and unity at the respective points of the prior distribution. It is however not the case that the logical operator – the implication $\Rightarrow$ – has the truth-value zero. It does not exist.

This is our normal condition of proof. We can neither show that $\mathbf{I}_\alpha \Rightarrow \alpha(G_0)$ nor that $\mathbf{I}_\alpha \Rightarrow \alpha(G_1)$. The conclusion that logical arguments are insufficient to determine a particular prior distribution lies close at hand. That is to say that a logically based information function ϕ simply does not exist in ”normal cases” (which are the cases who may be of any interest). I shall now argue that such a function still *can* be defined logically.

The foregoing argument shows that the concept of knowledge must be extended from just accommodating the collective pool of facts to embrace its logical consequences as well. If a ”new” fact follows logically from other, already accessible facts, we would know this new fact too, thanks to our knowledge of the former and the logical operators.

Assume two statement of fact F_0 and F_1. Also assume that $F_0 \Rightarrow F_1$. If F_0 is a fact F^*_0 , it follows that F_1 is also a fact F^*_1. If we both know F^*_0 and the relation $F_0 \Rightarrow F_1$, we also know that F^*_1 as soon as we carry out the logical operation $\Rightarrow$.

For us to *know* that F^*_1, without having any other sources to this knowledge than F^*_0 and the relation $F_0 \Rightarrow F_1$, three things are required: (1) that F^*_0 is a *known* fact, (2) that the relation $F_0 \Rightarrow F_1$ is a *known* relation, and (3) that we apply the operator $\Rightarrow$.

It is also important to distinguish between the *truth value* of a logical operator, and what I shall call the *knowledge value* of a logical operation.

The implication operator $\Rightarrow$ has the *truth value matrix*

	F_1 true (F^*_1)	F_1 false
F_0 true (F^*_0)	**True (1)**	**False (0)**
F_0 false	**True (1)**	**True (1)**

Table 1: The truth value matrix for logical implications.

Assuming that we always know how to apply the implication operator $\Rightarrow$, the following *knowledge value matrix* holds true

	F^*_1 known	F^*_1 unknown
$F_0 \Rightarrow F_1$ known	**Known (1)**	**Unknown (0)**
$F_0 \Rightarrow F_1$ unknown	**Unknown (0)**	**Unknown (0)**

Table 2: The knowledge value matrix for logical implications.

The interesting thing in our context is that all cells in the knowledge value matrix have zero value when the logical implication (the operator) is unknown. This is shown by the second row of the matrix, table 2.

There are good reasons to presume that a logical operator of the type ”partial implication” (RPI) also has this property. That is to say, if we do not know the relation, the knowledge value will be ”unknown” (or zero) even for the RPI as a whole. Since we do not know of any RPI’s in reality, it follows that in all situations where a RPI possibly *could be occurring*, the knowledge value would still be zero.

The point of a prior distribution is that it shall ”encapsulate” our prior knowledge – the prior information. But if we apply a logical RPI approach it must hold true that the ”encapsulated” prior knowledge are non-existent *even if we have access to a large quantity of prior information*. The reason for that is that we *do not know* the logical relation (whether it be an RPI or an ordinary two-valued implication) between the prior information and the prior distribution. From this one is tempted to conclude that an information function ϕ cannot be defined using the logical approach. But let us not give up!

What the ϕ-function is to achieve, is only a transformation of the prior information $\mathbf{I}_\alpha$ to a distribution of Kolmogorov weights $\alpha(G)$, i.e. to numerical values between zero and unity reflecting our prior knowledge with respect to different generator hypotheses. What, then do we *know* a priori about the generator hypotheses? Well, in the ”normal case” we *know* nothing! We surely *believe* this and that, but strictly speaking we *know* nothing, despite all conceivable prior information. The knowledge mass about the generator hypotheses following from the prior information is – sad to say – hopelessly non-existent.

We must now recall that the prior distribution in no way shows our entire knowledge mass. The only thing it shows is how much we know about the one hypothesis G_0 *in relation to* the other hypothesis G_1.[32]

This ”relation” is usually regarded as a *quotient* when talking about probabilities. If the probability of A is 0.25, and the probability of B is 0.75, then B is three times ”more probable” that A. Thus the quotient of the probabilities is three. But we could just as well say that B is 0.5 Kolmogorov units ”more probable” than A. Then the ”relation” between the probabilities is a difference.

Why this talk about ”relations”? Well, suppose the probability of A in the example were zero, and the probability of B were zero as well, then it makes a huge difference whether we talk about quotients of differences. The quotient between zero and zero is not defined, but the difference between zero and zero definitely is.

Let us briskly return to the prior distribution. It shows the quantitative ”relation” between our knowledge about the one hypothesis G_0 and the other hypothesis G_1. If this ”relation” refers to a quotient, the situation is bad. Since we do not know *anything* about any of G_0 or G_1, the quotient of these knowledge masses would be precisely ”zero divided by zero”, which is undefined.

But if the ”relation” refers to the difference between our knowledge about G_0 and G_1, then the whole thing turns out differently. We do neither know anything about G_0 nor about G_1, so the knowledge masses – be they non-existent – are equally large. The difference between them is clearly zero.

From this follows the only possible prior distribution. For if the prior distribution is to give equal Q-weights to all generator hypotheses (which mirror our knowledge position which is equally non-existent with regard to all hypotheses), as well as sum (integrate in the continuous case) to unity, then the prior distribution must be a rectangular distribution.

[32] Of course we are dealing with a number of generator hypotheses G, a number which does not even need be finite (in the continuous case the number of G-values is infinite). For the sake of reasoning we simplify by using only two G-values.

Thus, the prior distribution can be uniquely determined by using a logical approach, and that approach always follows (in the ”normal case”) the same line of argument and always yields the same result:

(1.) No *known* logical relation exists between the prior information and our knowledge about different generator hypotheses.
(2.) Therefore we *know* nothing a priori about any generator hypothesis.
(3.) Since our knowledge position is equally poor with regard to all generator hypotheses, the *difference* in knowledge mass as to the different hypotheses is zero.
(4.) To reflect this, the prior distribution must be rectangular.

Our knowledge position concerning the generator hypotheses may, despite all accessible prior information, be described as an empty board – a *tabula rasa*. Therefore we shall call the applicable prior distribution the *tabula rasa distribution* $\tau(G)$, which is uniformly distributed for all values of G.

(19.1.) $\tau(G) \in Re(0, 1)$. [33]

When we ask ourselves the simple question ”What is the aleatory probability $\Pi(A)$ of the present aleatory rigging in Ω?”, *de facto* we cannot refer to anything outside this rigging. Historical facts and experience give us no logical reason at all to favour the hypothesis G_0 and discriminate the hypothesis G_1. We must *unconditionally* ask for $\Pi(A)$ in Ω, we have no logically motivated right to favour or discriminate hypotheses. No matter how many facts we include in our prior information, our prior knowledge regarding $\alpha(G)$ is still a tabula rasa.

Thus the information function is extraordinarily simple. It yields the same prior distribution – the tabula rasa distribution $\tau(G)$ – regardless of the contents of the prior information $\mathbf{I}_\alpha$.

(19.2.) $\tau(G) = \phi(\mathbf{I}_\alpha)$

The relation ϕ between the prior information and the prior distribution is purely logical, involving no subjective or ”personal” judgement. Paradoxically, it is the circumstance of the prior information not implying anything about the generator hypotheses which enables us to draw the logical conclusion that the information function must always transform the prior

[33] This is the prior distribution used by Laplace, and – some say – Bayes himself. The principle is to apply a rectangular prior distribution when prior information is lacking is often called the ”principle of non-sufficient reason”, or the ”principle of indifference”, as Keynes names it. *Vide* Keynes (1921), chapter 4 (”The Principle of Indifference”), and Hacking (1975), chapter 14 (”Equipossibility”) for a thorough discussion and further references.

information to one and only one kind of prior distribution, namely the tabula rasa distribution. ”The choice of prior distribution” is therefore a somewhat misleading expression, for the prior distribution chooses itself, by pure logic.

20. Epistemic probability and epistemic weight

According to Ian Hacking, the modern concept of probability emerged in the 1660’s.[34] The word probability is much older that that, as are many other, more primitive notions of the probability concept. Since the 17:th Century a vast number of improvement have been made in the technique of probability calculus and inferences. The interesting point in our connection is however that both theses aspects – prior calculus and inferences – were there from the genesis of the probability concept. Hacking emphasises that the probability concept always (or at least during its modern existence) been ”two-sided” or dual. On the one side there has been ”aleatory” probability, on the other side ”epistemic” probability.

It is a fact that we do not have, and in many cases never will have, exact quantitative knowledge of the true aleatory probability of empirical phenomena. This state of affairs implies a need for epistemic concepts, by the aid of which we may characterise our knowledge position about the aleatory probability we seek. Keynes argues emphatically, and very rightly, that such an epistemic concept must be two-dimensional.

Keynes called the one dimension the ”probability of the argument” and the other the ”weight of the argument”. We shall adopt Keynes’s categorisation in probability and weight, respectively, but stress that we give those concepts meanings similar, but not equal, to what Keynes did. We shall also add the adjective ”epistemic” to distinguish these concepts from the ”aleatory” counterpart. What, then, do epistemic probability and epistemic weight really mean?

We have already discussed the concept of aleatory probability in some detail, but the meaning of the concept of epistemic probability was only briefly hinted at. It is really rather regrettable that we could not proceed to the definition of the concept of epistemic probability until now. But the thing is that the long journey here has been altogether necessary to clear up the many enclosing problems.

As was mentioned initially, the word ”episteme” means eternal and unchanging knowledge, the acquisition of which is one of the three virtues of Aristotle’s *Nichomachian Ethics*. While aleatory probability refers to the *rerum natura*, the nature of things – the propensities of events to occur – the epistemic probability refers to our *knowledge about the propensities of events to occur*.

Aleatory probability is fundamentally objective. It is a property of nature that we cannot change. According to Hacking, epistemic probability is fundamentally subjective. It is about

[34] *Vide* Hacking (1975), but also his exquisite *Taming of Chance* (1991).

what we know, and what we know, Hacking tacitly argues, dwells in the minds of people and therefore it must be subjective. I remain sceptical to this train of thought.

Hacking’s view that in the end, epistemic probability is subjective, appears erroneous to me. The notion of anything being *episteme* – eternal, universal and true knowledge – also being subjective, appears a self-contradiction to me. If there is such a thing as epistemic probability, then surely it is not subjective, I would say. The Eternal, the Universal and the True must reasonably be objective, and so must epistemic probability.

Epistemic probability reminds of what is usually called logical probability. Epistemic probability is logical in the sense that it expresses an objective relation between a set of evidence from an aleatory rigging, and an *empirical hypothesis* about the kind of event defined in that rigging.

Recall that an empirical hypothesis *H* is a proposition of the kind ”*a* will be the case”, where *a* is a future event, or ”*a* was the case”, where *a* is an unknown fact. *H* does *not* denote a variable, but a ”fixed” proposition. First we distinguished between the empirical–future hypotheses and the empirical–historical hypotheses, and found that the former type can be neither true nor false, but that the latter type must be either true or false. Then we concluded that if we are dealing with *unknown* facts, we could reason ”as if” empirical–historical hypotheses were propositions about future events.

A hypothesis is never aleatorily probable, that goes for empirical hypotheses too. But empirical hypotheses are epistemically probable. This is unique to empirical hypotheses – generator hypotheses are *not* epistemically probable.

The events that empirical hypotheses *H* makes such categorical statements about, are either unknown facts or future events (which have not yet occurred and which we do not know whether they will occur or not). At first sight, the categorical formulation may appear somewhat odd, but at closer inspection, it is fully reasoned. Just think of the opposite, that we would formulate an empirical–future hypothesis *H*: ”*a might* occur”. How can we state whether such a hypothesis is epistemically probable? We cannot. The moment of uncertainty (the ”might” moment) in this example lies *within* the proposition. It must be moved out of the proposition to enable us to speak *about* the proposition as epistemically probable.

Epistemic probabilities obey the Kolmogorovian axioms – they are Kolmogorov weights, to which we add a particular philosophical interpretation. Strictly speaking an epistemic probability is always a *conditional* Kolmogorov weight. The conditioning refers to the evidence presented, preferably expressed by the value of the evidence function *E*.

The presupposition for an empirical hypothesis *H:* ”*a* will be (was) the case” to be epistemically probable with regard to the evidence *E*, is that the event *a* of type *A* is spawned by an aleatory rigging Π(*A*) in the event space Ω.

The *epistemic probability of H, given E,* we shall denote by $P(H \mid E)$, and it is mathematically defined as the expected value of the Bayesian posterior distribution E[ζ] when the tabula rasa distribution τ(*G*) is applied as prior.

(20.1.) $P(H \mid E) = \mathrm{E}[\zeta \mid \alpha = \tau]$

Nota bene this is only the *mathematical* definition. The epistemic probability $P(H \mid E)$ always numerically coincides with what we previously called a ”unweighted WAL estimate” of the aleatory probability Π(*A*). But, as we recall, the ”unweighted WAL estimate” did not necessarily have any particular philosophical interpretation. It is only in interpretation that the epistemic probability $P(H \mid E)$ differs from an ” unweighted WAL estimate”. The epistemic probability is an objective numerical expression for *what we know* about the aleatory probability Π(*A*) in Ω.

For every epistemically probable hypothesis *H*, there also exists a number $W(E \mid H)$ – *the epistemic weight of the evidence E, with respect to the empirical hypothesis H* – which is mathematically defined as the standard deviation of the Bayesian posterior distribution $\mathrm{D}[\zeta \mid \alpha = \tau]$, adjusted by multiplication by the inverted value $\mathrm{D}^{-1}[\alpha]$ of the standard deviation of the tabula rasa distribution.

(20.2.) $W(E \mid H) = 1 - \mathrm{D}^{-1}[\tau] \cdot \mathrm{D}[\zeta \mid \alpha = \tau]$

The epistemic weight $W(E \mid H)$ always numerically coincides with what we previously called the ”weight of the evidence with respect to (the unweighted) WAL estimate” of Π(*A*) in Ω. But that magnitude did not necessarily have any particular philosophical interpretation. It is only in the interpretation that the epistemic weight $W(E \mid H)$ differs from ”the weight of the evidence...etc.”. The epistemic weight is an objective numerical expression for *how much we know* about the aleatory probability Π(*A*) in Ω.

Epistemic probability is nothing but the epistemic correspondent to aleatory probability. Aleatory probability refers to ”the propensity to occur”, and epistemic probability to ”what we know about the propensity to occur”. Alternatively, epistemic probability could be expressed as the ”objective location estimate” of an aleatory probability.

Epistemic weight does not correspond to the location, but to the degree of precision, [35] of our estimate of the true aleatory probability. Obviously, a well-founded estimate, which is built on a large volume of evidence, must have a larger precision than a poorly founded estimate, built on scanty evidence. In a way, the weight of evidence may be regarded an ”objective indicator” of the quantity of evidence, or *how much we know* about ”the propensity to occur”. We might also speak of the weight of evidence as an expression of the balance between knowledge and ignorance, or the balance between what we do, and do not, know.

It is important to always characterise our knowledge position by stating *both* the epistemic probability of *H*, given *E*, *and* the epistemic weight of the evidence *E* with respect to *H*. *Thus we should always state a pair of numbers* $[P(H|E)\ ,\ W(E|H)]$ *and not only one of the two numbers*. Lest we do, only one dimension will be reflected of our two-dimensional epistemic position.

In the end, the concepts of epistemic probability and epistemic weight are not very complicated. To comprehend their meaning must be an intuitive process. To compute them numerically is not very complicated either, much thanks to the tabula rasa distribution being generally applicable as prior.

Indeed, we only have two cases, which we shall call *the discrete case* and *the continuous case*, respectively These cases each correspond to one type of event space. The discrete case is strictly applicable when the event space Ω accommodates a finite number of *A*-premises (when the population is finite); the continuous case applies when the number of *A*-premises is infinitely large (the population is unlimited). In practice we disregard vi the population being finite, provided it is large enough to make computations from the assumption of an infinite population good numerical approximations.

For each case we have one, and only one, definite mathematical formula for the computation of the epistemic probability, and one, and only one, definite mathematical formula for the computation of the epistemic weight. Thus, there are four formulas all in all. These formulas are to be found in appendices I (the continuous case) and II (the discrete case), respectively.

[35] ”Precision” should be understood in a loose sense. We are not talking about the measure which is usually called precision, and which is defined as the inverted value of the variance V^{-1}.

Epistemic probability and weight are uniquely determined by the evidence [*E*, *m*, *n*] at hand. In the continuous case, however, the size of the population *m* is omitted. The fact that only three parameters affect the computations makes possible and desirable to ”once and for all” carry out thorough numerical computations of epistemic probabilities and weights for a large number of evidence arrays [*E*, *m*, *n*], and to cross tabulate the results like this is usually done in tables of statistical distributions.

In particular, the continuous case ought to be manageable, considering that only two parameters [*E*, *n*] are involved. The tabulation of the continuous case will thus be two-dimensional (in the same fashion as the binomial distribution). The discrete case, however, requires three dimensions (in the same way as the F-distribution), and will therefore me more space-consuming as one cross-tabulation will be needed for every population size *m*.

As the sample (or the population) grows, the formulas will contain very large numbers. For this reason, manual computation is unthinkable. Computer assistance is a necessity. My time frames for writing this essay have unfortunately not been generous enough to allow the required programming and computation. These computations are an urgent future task.

21. Appendix I: The continuous case

We have the evidence [*E*, *n*] from this rigging, and we would like to draw conclusions about *the epistemic probability P*(*H*|*E*). In the continuous case, we are dealing with an aleatory rigging Π(*A*) in Ω. The population is infinite ($m \to \infty$). Then, it is true that the likelihood function obeys a binomial distribution in *E* and *n*.

$$(21.1.)\qquad \lambda(G, E, n) = \pi(E|G) \in Bin(n, E) = \begin{bmatrix} n \\ nE \end{bmatrix} \cdot G^{nE} \cdot (1 - G)^{n-nE}$$

The tabula rasa distributions is given by

$$(21.2.)\qquad \tau(G) \in Re(n, E) = 1\ ,$$

which yields the posterior distribution

$$(21.3.)\qquad \zeta(G|E) = \frac{\tau(G) \cdot \pi(E|G)}{\int \tau(G) \cdot \pi(E|G)\, \mathrm{d}G} = (n{+}1) \cdot \begin{bmatrix} n \\ nE \end{bmatrix} \cdot G^{nE} \cdot (1 - G)^{n-nE}$$

The epistemic probability of the empirical hypothesis *H*, given the evidence *E*, is given by the expected value of the posterior distribution E[ζ]

$$(21.4.)\qquad P(H|E) = \mathrm{E}[\zeta(G|E)] = \int G \cdot (n{+}1) \cdot \begin{bmatrix} n \\ nE \end{bmatrix} \cdot G^{nE} \cdot (1 - G)^{n-nE}\, \mathrm{d}G$$

where the integral runs from zero from unity. By binomial expansion and integration we obtain

$$(21.5.)\qquad P(H|E) = (n{+}1) \cdot \begin{bmatrix} n \\ nE \end{bmatrix} \cdot \sum \begin{bmatrix} n-nE \\ k \end{bmatrix} \cdot (-1)^{k+2}/(k{+}2)$$

where the sum runs from $k = 0$ to $n{+}nE$.

The epistemic weight of the evidence *E*, with respect to the empirical hypothesis *H*, is given by

$$(21.6.)\qquad W(E|H) = 1 - \mathrm{D}^{-1}[\tau] \cdot \mathrm{D}[\zeta] = 1 - (1/\sqrt{12}) \cdot \mathrm{D}[\zeta]$$

because D[τ] = 1/√12. D[•] expresses the standard deviation of the distribution in question. We start out by computing V[ζ], the square root of which is D[ζ]. It is true that V[ζ] = E[ζ²] – E²[ζ]. E[ζ²] is given by

$$(21.7.)\qquad \mathrm{E}[\zeta^2] \;=\; \int G^2 \cdot (n{+}1) \cdot \left[\begin{matrix} n \\ nE \end{matrix}\right] \cdot G^{nE} \cdot (1 - G)^{n-nE}\,\mathrm{d}G$$

where the integral runs from zero to unity. Binomial expansion and integration yields

$$(21.8.)\qquad \mathrm{E}[\zeta^2] \;=\; (n{+}1) \cdot \left[\begin{matrix} n \\ nE \end{matrix}\right] \cdot \sum \left[\begin{matrix} n{-}nE \\ k \end{matrix}\right] \cdot (-1)^{k+3}/(-k{-}3)$$

where the sum runs from $k = 0$ to $n{+}nE$.

Thus, we obtain for the variance V[ζ],

$$(21.9.)\qquad \mathrm{V}[\zeta] \;=\; \mathrm{E}[\zeta^2] - \mathrm{E}^2[\zeta] \;=\; \Big[(n{+}1) \cdot \left[\begin{matrix} n \\ nE \end{matrix}\right] \cdot \sum \left[\begin{matrix} n{-}nE \\ k \end{matrix}\right] \cdot (-1)^{k+3}/(-k{-}3)\Big] -$$

$$- \;\Big[(n{+}1) \cdot \left[\begin{matrix} n \\ nE \end{matrix}\right] \cdot \sum \left[\begin{matrix} n{-}nE \\ k \end{matrix}\right] \cdot (-1)^{k+2}/(k{+}2)\Big]^2$$

where both sums run from $k = 0$ to $n{+}nE$. The standard deviation D[ζ] = √V[ζ] is inserted into (6.) above, whereby the epistemic weight of the evidence $W(E\,|\,H)$ obtains.

22. Appendix II: The discrete case

We have the evidence $[E, m, n]$ from this rigging, and we would first like to draw conclusions about *the epistemic probability* $P(H \mid E)$. In the discrete case we are dealing with an aleatory rigging $\Pi(A)$ in Ω. The population is finite (m is ”small”). Then, it is true that the likelihood function obeys a hypergeometric distribution in E, m and n.

$$(22.1.) \qquad \Lambda(G_i, E, m, n) \; = \; \Pi(E \mid G_i) \; = \; \frac{\begin{bmatrix} G_i \\ nE \end{bmatrix}\begin{bmatrix} m - G_i \\ n - nE \end{bmatrix}}{\begin{bmatrix} m \\ n \end{bmatrix}}$$

where G_i refers to the i:th of the (m+1) possible generator hypotheses. In the discrete case, G is a discrete variable, which can assume $r = (m+1)$ different values, where $m \geq 1$ is a natural number. Hence, $r \geq 2$.

$$(22.2.) \qquad G \; = \; [G_1, G_2, \ldots , G_{r-1}, G_r] \; = \; [0/m, 1/m, \ldots , ((m-1)/m), m/m]$$

The tabula rasa distribution $\tau(G)$ is given by

$$(22.3.) \qquad Q(G_i) \; = \; 1/r \, , \qquad i = 1, 2, \ldots , r.$$

which yields the posterior distribution

$$(22.4.) \qquad \zeta(G_i \mid E) \; = \; \frac{Q(G_i) \cdot \Pi(E \mid G_i)}{\sum Q(G_i) \cdot \Pi(E \mid G_i)} \; = \; \Psi / \sum \Psi$$

where $\Psi \; = \; 1/r \cdot \begin{bmatrix} G_i \\ nE \end{bmatrix}\begin{bmatrix} m - G_i \\ n - nE \end{bmatrix}$, and where the sum runs from $i = 1$ to $r = m+1$.

The epistemic probability of the empirical hypothesis H, given the evidence E, is given by the expected value of the posterior distribution $\mathrm{E}[\zeta]$

$$(22.5.) \qquad P(H \mid E) \; = \; \mathrm{E}[\zeta(G_i \mid E)] \; = \; (1/ \sum \Psi) \cdot \sum G_i \cdot \Psi$$

where the sums run from $i = 2$ to r.

The epistemic weight of the evidence E, with respect to the empirical hypothesis H, is given by

(22.6.) $$W(E \mid H) = 1 - \mathrm{D}^{-1}[\tau] \cdot \mathrm{D}[\zeta]$$

The standard deviation of the tabula rasa distribution is not constant in the discrete case, but depending on the population size m. The variance V[τ] is given by

(22.7.) $$\mathrm{V}[\tau] = \mathrm{E}[G^2] - \mathrm{E}^2[G] = (1/r) \sum G_i^2 - [\sum G_i/r]^2 = [\sum i/r - (\sum i)^2]/r^2$$

where the sums run from $i = 2$ to r. This variance declines with an increasing population size (and, thereby, an increasing r). The maximum value is 1⁄4 when $m = 1$, and the value decreases, rapidly for a start, then slower, when m is increased. The variance V[τ] converges to 1⁄12 as $m \rightarrow \infty$, i.e. when the discrete case approaches the continuous.

Thus, in the discrete case we must compute the adjustment factor $\mathrm{D}^{-1}[\tau]$ from time to time, in order to obtain the epistemic weight of the evidence with respect to the hypothesis H. But besides this, the procedure is analogous to the continuous case. We compute V[ζ], the square root of which is D[ζ]. It is true that $\mathrm{V}[\zeta] = \mathrm{E}[\zeta^2] - \mathrm{E}^2[\zeta]$. $\mathrm{E}[\zeta^2]$ is given by

(22.8.) $$\mathrm{E}[\zeta^2] = (1/\sum \Psi) \cdot \sum G_i^2 \cdot \Psi$$

where the sums run from $i = 2$ to r.

Thus, we obtain for the variance V[ζ],

(22.9.) $$\mathrm{V}[\zeta] = \mathrm{E}[\zeta^2] - \mathrm{E}^2[\zeta] = \Big[(1/\sum \Psi) \cdot \sum G_i^2 \cdot \Psi \Big] - \Big[(1/\sum \Psi) \cdot \sum G_i \cdot \Psi \Big]^2$$

where both sums run from $i = 2$ to r. The standard deviation $\mathrm{D}[\zeta] = \sqrt{}\mathrm{V}[\zeta]$ is inserted into (6.) above, whereby the epistemic weight of the evidence $W(E \mid H)$ is obtained.

References

Bergström, Ingmar and Forsling, Wilhelm (1992): *I Demokritos' fotspår* [”In the footsteps of Democrit”] (Natur & Kultur).

Blom, Gunnar (1980): *Sannolikhetsteori och statistikteori med tillämpningar* [”Probability theory and statistical theory with applications”] (Studentlitteratur).

Carnap, Rudolf (1950): *The Logical Foundations of Probability* (University of Chicago Press).

de Finetti, Bruno (1972): *Probability, Induction and Statistics – The Art of Guessing* (Wiley).

de Finetti, Bruno (1990): *Theory of Probability* (2 volumes, Wiley).

Fisher, R.A, (1925): *Statistical Methods for Research Workers* (Oliver & Boyd).

Flyvbjerg, Bent (1994), ”Vårt behov av Phronesis” [”Our need for Phronesis”], *Ordfront Magasin* 3.

de Groot, Morris H. (1970): *Optimal Statistical Decisions* (McGraw Hill).

Hacking, Ian (1965): *The Logic of Statistical Inference* (Cambridge University Press).

Hacking, Ian (1975): *The Emergence of Probability* (Cambridge University Press).

Hacking, Ian (1991): *The Taming of Chance* (Cambridge University Press).

Hawking, Stephen W. (1989): *A Brief History of Time* (Swedish translation: *Kosmos*, Prisma Magnum, 1992).

Hogg, R.V. and Tanis, E.A. (1983): *Probability and Statistical Inference* (3:rd edition, Macmillan).

Hume, David (1748): *An Inquiry concerning the Human Understanding*.

Keynes, John Maynard (1921): *Treatise on Probability* (Macmillan; reprinted as volume VIII of *The Collected Writings of John Maynard Keynes*, Macmillan, 1973–1989).

Keynes, John Maynard (1938): ”My Early Beliefs”, in *Two Memoirs* (first edition 1949, reprinted in *Essays in Biography,* volume X of *The Collected Writings of John Maynard Keynes* (Macmillan 1973–1989).

Kolmogorov, A.N. (1933): *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Springer Verlag). English translation: *Foundations of the Theory of Probability* (Chelsea Publishing Co, 1950).

de Laplace, Pierre Simon (1814): *Théorie analytique des probabilités* (2:nd edition, Paris).

Moore, George Edward (1903): *Principia Ethica* (Cambridge University Press).

Ramsey, Frank P. (1926, 1928): ”Truth and Probability” and ”Further Comments” in *The Foundations of Mathematics and Other Logical Essays,* (Kegan Paul).

Runde, Jochen (1990): ”Keynesian Uncertainty and the Weight of Arguments”, *Economics and Philosophy,* 6, 275–292.

Runde, Jochen (1991): ”Keynesian Uncertainty and the Instability of Beliefs”, *Review of Political Economy*, 3.2, 125–145.

Savage, Leonard J. (1954): *The Foundations of Statistics* (2:nd edition, Dover Publishers, 1972).

Shackle, G.L.S. (1974): *Keynesian Kaleidics* (Edinburgh University Press).