EuJAP | Vol. 10 | No. 1 | 2014
UDK 165.1/.2
519.2
1 Bayes, T.
53-05 Cox, R. T

WHAT IS PROBABILTY AND WHY DOES IT MATTER

ZVONIMIR ŠIKIĆ
University of Zagreb

ABSTRACT

The idea that probability is a degree of rational belief seemed too vague for a foundation of a mathematical theory. It was certainly not obvious that degrees of rational belief had to be governed by the probability axioms as used by Laplace and other prestatistical probabilityst. The axioms seemed arbitrary in their interpretation. To eliminate the arbitrariness, the statisticians of the early 20th century drastically restricted the possible applications of the probability theory, by insisting that probabilities had to be interpreted as relative frequencies, which obviously satisfied the probability axioms, and so the arbitrariness was removed. But the frequentist approach turned more subjective than the prestatistical approach, because the identifications of outcome spaces, the choices of test statistics, the declarations of what rejection regions are, the choices of null-hypothesis among alternatives, the contradictory choices between sizes and powers etc., depend on thoughts or even whims of the experimenter. Frequentists thus failed to solve the problems that motivated their approach, they even exacerbated them. The subjective Bayesianism of Ramsey and de Finetti did not solve the problems either. Finally Cox provided the missing foundation for probability as a degree of rational belief, which makes the Bayesian probability theory (which is based on this foundation) the best theory of probable inference we have. Hence, it is quite unbelievable that it is not even mentioned in recent philosophy textbooks devoted to the probable inference. The reason could be that it requires fairly sophisticated mathematics. But not even to mention it? We explain the history and prove Cox theorem in a novel way.

Keywords: probability, subjective Bayesianism, logical Bayesianism, Cox theorem

Probability has a mathematical aspect and a scientific aspect. There is a reasonable agreement about the mathematics of probability. Almost everybody accepts the same probability axioms and has no disputes about the truths of the mathematical theory of probability. Yet, when it comes to scientific applications of the theory there are different opinions about what probabilities are. Some identify them with degrees of (rational) belief, some with limiting frequencies, and there are other opinions. Why does it matter?[1] We explain why, starting with a simple problem of a coin fairness testing.[2]

Consider a hypothesis about the probability of a coin coming up heads. If we denote it by H, then $H = 0$ and $H = 1$ represent a coin which, respectively, produces a tail or a head on every flip. There is a continuum of possibilities between these extremes, with $H = 1/2$ indicating a fair coin. Now, if you had observed 3 heads in 12 flips, do you think it was a fair coin?

[1] Think about mathematics of numbers, i.e. arithmetic. In scientific applications it does not matter what numbers are. What is important are rules that numbers obey, not what they really are. Is it not the same with probabilities?

[2] The idea is to compare Laplacean (prestatistical) approach to a textbook problem, with the statistical approach to the same problem. It may seem that this is more appropriate for a college course than for a research article, but my experience is that such an introduction is eye opening (and surprising) even for the audiences that are highly trained in probability. The problem is taken from Sivia 1996, ch.2.

A Bayesian who thinks of probabilities as degrees of rational belief, will use Bayes' theorem to answer this question:[3]

$$pr(H \mid D, I)dH = \frac{pr(H|I)dH \cdot pr(D|H,I)}{pr(\mathrm{D}|I)}$$

Actually, he will use a simpler form:

$$pr(H \mid D,I) \propto pr(H \mid I)\, pr(D \mid H, I),$$

because he can evaluate the missing constant (which does not depend on H) from the normalisation condition

$$\int_0^1 pr(H|D,I)dH = 1.$$

The power of the theorem lies in the fact that it relates the probability that the hypothesis H is true, given the data D (e.g. 3 heads in 12 flips) and background information I (e.g. flipping is vigorous, coin is symmetric etc.), to the probability that the data would have been observed if the hypothesis is true, which is easier to assign.

Prior probability $pr(H \mid I)$ represents the degree of rational belief in H given I (with no data D available). It is modified by the data D, through the *likelihood* $pr(H \mid D, I)$, and yields the *posterior* probability $pr(H \mid D, I)$, which represents the degree of rational belief in H given I and the data D.

In our specific case of coin flipping, *prior* $pr(H \mid I)$ represents what is known about the coin before any data is taken into account. The state of ignorance is represented by the uniform probability assignment

$$pr(H \mid I) = \begin{cases} 1 & 0 \leq H < 1 \\ 0 & \text{otherwise} \end{cases}$$

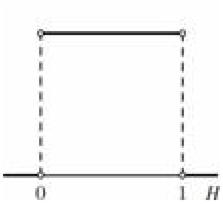

Fig. 1

This prior state of ignorance is modified by the data through *likelihood*:

$$pr(D \mid H, I) \propto H^R(1-H)^{N-R},$$

[3] In what follows *pr* is probability or probability density which, for short, we also call probability.

where H is the probability of obtaining a head and R is a number of heads obtained in N flips. (For simplicity, equality is again replaced by proportionality, since the omitted term does not depend on H.) By Bayes' theorem:

$$pr(D \mid H, I) \propto H^R(1-H)^{N-R}, \text{ for } 0 \leq \mathrm{H} \leq 1,$$

otherwise it is 0. If the coin is flipped once and it comes up heads, the resulting posterior is:[4]

$$pr(H \mid \{h\}, I) \propto H$$

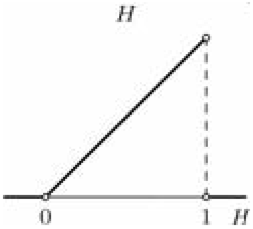

Fig. 2.

If the coin is flipped for a second time and again comes up heads, the resulting posterior is:

$$pr(H \mid \{h,h\}, I) \propto H^2$$

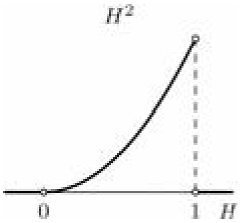

Fig. 3.

If the third flip comes up tails the resulting posterior is:

$$pr(H \mid \{h,h,t\}, I) \propto H^2(1\text{-}H)$$

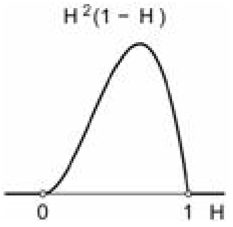

Fig 4.

If the forth flip also comes up tails the resulting posterior is:

[4] The graphs are normalized in such a way that the maxima are always 1.

$pr(H \mid \{h,h,t,t\}, I) \propto H^2(1\text{-}H)^2$

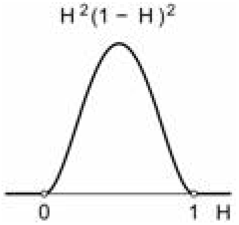

Fig. 5.

And so on. The following figures show how the posterior evolves as the number of data becomes larger and larger.[5] The position of the maximum wobbles around, but the wobbling decreases with the increasing amount of data. The width of the posteriors also becomes narrower with more data. For the coin in our example, the best estimate of *H* converges to 0.25 (of course this was the value used to simulate the flips).

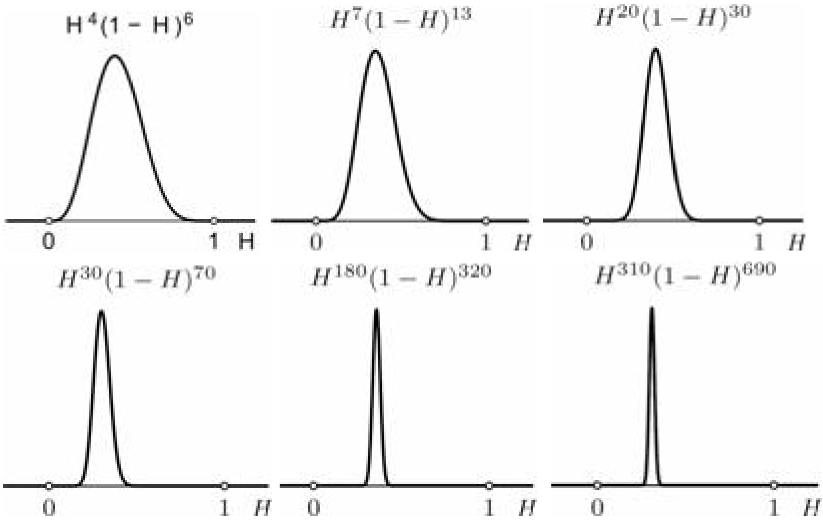

Fig. 6.

People tend to agree with the binomial distribution for the likelihood ($pr(D \mid H, I) \propto H^R(1–H)^{N\text{-}R}$) but worry about the prior: how would the inference about the coin have changed if a different prior was chosen? If we repeated the analysis of the date with different priors the results would have been the same, because the posterior is dominated by the likelihood, and the choice of the prior is largely irrelevant (cf. fig. 2.2. in Sivia 1996).

[5] It is easy to prove that it does not matter whether the data are introduced one by one or all of them simultaneously.

The result of the Bayesian analysis is the probability distribution of every possible hypothesis H, given one real data set D.

Often, we wish to summarize this distribution with just two numbers: the best estimate and its reliability. If we denote the posterior by P, i.e. $P = pr(H \mid D, I)$, then the best estimate of its value is the maximum point H_0 given by:

$$\frac{dP}{dH} \Big|_{H_0} = 0 \quad \text{and} \quad \frac{d^2P}{dH^2} \Big|_{H_0} < 0.$$

The measure of the reliability of the best estimate is the spread of the posterior about it. The behaviour of any function around a point can be estimated by its Taylor's expansion about the point. But, rather than dealing with posterior P, it is easier to deal with its logarithm $L = \ln P$. Expanding L about H_0, we get:

$$L \approx L(H_0) + \frac{1}{2}\frac{d^2L}{dH^2} \Big|_{H_0} (H - H_0)^2 ,$$

(the linear term is 0 because L is monotone in P, so H_0 as the maximum point of P is also the maximum point of L).

If we define $1/\sigma^2 = -(d^2L/ dH^2) |_{\mathrm{H0}}$ we get:

$$L \approx L(H_0) - \frac{(H - H_0)^2}{2\sigma^2}$$

which by exponentiation yields:

$$P = pr(H \mid D, I) \approx P(H_0)\, e^{-\frac{(H - H_0)^2}{2\sigma^2}}$$

From the normalization condition:

$$1 = \int_0^1 pr(H|D,I)dH \approx \int_0^1 P(H_0)\, e^{-\frac{(H - H_0)^2}{2\sigma^2}}\, dH = P(H_0)\sigma\sqrt{2\pi}$$

it follows that $P(H_0) = 1/\sigma\sqrt{2\pi}$, i. e.

$$pr(H \mid D, I) \approx \frac{1}{1/\sigma\sqrt{2\pi}} e^{-\frac{(H - H_0)^2}{2\sigma^2}}$$

This normal or Gaussian approximation[6] is usually conveyed by the statement:

$$H = H_0 \pm \sigma,$$

with H_0 the best estimate and σ referred to as the *error-bar*. The integral properties of the normal approximation tell us that the probability that the true value of H lies within ± σ of H_0 is 67% i. e.

$$pr(H-\sigma \leq H \leq H_0+\sigma|D, I\,) = \int_{H_0+\sigma}^{H_0+\sigma} pr(H|D,I)dH \approx 0{,}67$$

Similarly, the probability that H lies within ± 2σ of H_0 is 95%, and that H lies within ± 3σ of H_0 is 99.8%.

In the coin example:

$$P = pr(H \mid D, I\,) \propto H^R(1-H)^{N-R},\ 0 \leq \mathrm{H} \leq 1,$$
$$L = C + R \ln H + (N-R)\ln\,(1-H),$$

$$\frac{dL}{dH} = \frac{R}{H} - \frac{N-R}{1-H} = 0 \ \text{ for } \ H_0 = \frac{R}{N}\,,$$

$$\frac{d^2L}{dH^2}\Big|_{H_0} = -\frac{R}{H_0^2} - \frac{(N-R)}{(1-H_0)^2} = -\frac{N}{H_0(1-H_0)}\,,$$

$$\sigma = \sqrt{\frac{H_0(1-H_0)}{N}} < \frac{1}{\sqrt{N}}$$

Hence, the relative frequency of heads R/N is the best estimate of H, and its error-bar σ is less than $1\sqrt{N}$. So, the width of the posterior becomes narrower with the increasing number of the data N, as we have seen above (cf. fig. 6.).[7]

This prestatistical approach to our problem was the standard approach of Laplace and his contemporaries. As we have just seen, the approach is extremely successful. But nevertheless, it was rejected by the frequentists of the late 19th and the early 20th century. Why? The idea that probability is a degree of rational belief seemed too vague for a foundation of a mathematical theory. It was certainly not obvious that degrees of rational belief had to be governed by the probability axioms used by Laplace and others. The axioms seemed arbitrary in this interpretation.[8] To eliminate this arbitrariness, the mathematicians of the late

[6] The approximation is just the quadratic approximation of the logarithm and has nothing to do with coins or probabilities.

[7] The last formula also proves that it is easier to identify a highly biased coin than it is to be confidential that it is fair, because the nominator $H_0(1-H_0)$ is greatest when $H_0 = 1/2$.

[8] Furthermore, the probability rules described how to manipulate probabilities, but they did not specify how to assign the prior probabilities that were being manipulated. We will not address this

19th and the early 20th century drastically restricted the possible applications of the theory, by insisting that probabilities had to be interpreted as relative frequencies of occurrences in repeated random experiments (mostly imagined, rarely actual). The relative frequencies obviously satisfied the probability axioms, hence their arbitrariness was removed. Also, the frequentist approach, by its reference to observation of repeated experiments, seemed to make probability an objective property of "random phenomena" and not a subjective degree of the rational belief of Bayesians.

But, the frequency definition of probability made the concept of the probability of a hypothesis illegitimate, e. g. the prior $pr(H|I)$ and the posterior $pr(H|D, I)$ in the coin example make no sense. A hypothesis is either true or false, it is not a random variable. A consequence is that scientists are not allowed to use Bayes' theorem to asses hypotheses. So, how would a frequentist deal with the coin fairness problem? He can not calculate the probability of the fairness hypothesis (the hypothesis that $H = 1/2$), even less the probability distribution of every possible hypothesis H, given the data D, since hypotheses have no probabilities.

Hence, Fisher developed his system of significance tests for hypotheses testing. To perform the test, an experiment must be devised, in our example flipping the coin a predetermined number of times, say 12, and then the result analysed in three steps.

First, specify the outcome space. In our example 2^{12} possible sequences of 12 heads or tails. The result of the experiment should be summarised in some numerical form, e.g. the number of heads in the outcome. This summary is called *test-statistics*, and as a function of outcomes it is a random variable which has probability.

Second, calculate the probability of every possible value of the test-statistics, given the hypothesis you are testing (Fisher called it the *null-hypotheses*). This is the *sampling distribution* of the test-statistics. In our case it is $pr(R) = \binom{12}{R}(1/2)^R(1/2)^{12-R}$, with R the number of heads:

R	*pr R*	*R*	*pr R*	*R*	*pr R*	*R*	*pr R*
0	0.0001441406	3	0.053710938	6	0.225585938	9	0.053710938
1	0.0029296875	4	0.120849609	7	0.193359375	10	0.0161132813
2	0.0161132813	5	0.193359375	8	0.120849609	11	0.0029296875
						12	0.0002441406

Third, look at all results which *could have* occurred (given the null-hypothesis) and which,

question, although we have seen that at least in some cases, e. g. our coin example, it is irrelevant. Furthermore, probability is the logic of uncertainty, and as the standard logic does not tell us what are the factual truths, but only what follows from what, so probability does not tell us what are prior probabilities but only what probability follows from given probabilities.

as Fischer put it, are more extreme than the result that did occur. It means their probability is less than or equal to the probability of the actual outcome. Then calculate the probability *pr** that the outcome will fall within this group. For example, if our experiment produced 3 heads in 12 flips, the result with less or equal probabilities to this are *R* = 0,1,2,3,9,10,11,12; and the probability of at least one of them occurring (c.f. the shaded values in the table above) is *pr**= 0.15. Fisher's accepted convention is to *reject* the null-hypothesis just in case pr*≤ 0.05. Hence our null-hypothesis of the fairness of the coin is not rejected.

Some statisticians recommend 0.01 or even 0.001 as the critical *pr**. The adopted critical probability is called the *significance level* of the test, and the null-hypothesis is said to be *rejected at this significance* level if *pr** is less than or equal to it.

"The null-hypothesis is rejected at a significance level" is a technical expression, which means that the result of the experiment fall in a certain region (declared "the rejection region"). But what does it really say about the null-hypothesis? Today the standard view (introduced by Neyman) is that a rejection or non-rejection of a null-hypothesis is not an inductive inference, but just an instruction for inductive behaviour. If we behave according to the instruction, in the long run we shall reject a true hypothesis H, i.e. we shall make a type I error, no more than once in a hundred times, when significance level is 0.01.

We may also worry, as Neyman and Pearson did, about accepting a false hypothesis *H*, i.e. making a type II error. The probability of type II error is the probability of rejecting a true alternative hypothesis, let's call it H_a[9], by accepting the false *H*. The complement of the significance level of rejecting H_a is called the *power* of a test and, in this context, the significance level of rejecting *H* is called its *size*. An ideal would be to maximize the power and to minimize the size of a test. But that ideal is inconsistent. In most cases a contraction in size brings with it an expansion in power, and vice versa.

Apart from the volatility of what is declared to be "the rejection region", the incoherence of contracting the size and expanding the power of a test, and considering only one or two hypotheses[10], there are other problems with the frequentist approach.

For example, different random variables may by defined on an outcome space, not all of them leading to the same conclusion when used as a test-statistics in a significance test. This is the notorious problem of "which test-statistics to choose?"

There is also the problem of "the stopping rule". Consider again that a coin has been flipped 12 times, giving 3 heads and 9 tails. Is this the evidence that the coin is biased? With the data provided, the frequentists cannot even begin to answer this question. Namely, from these data it is not clear what the outcome space for the data is. If the frequentist is told that the experimenter's plan was to flip the coin 12 times, then analysis can proceed as above. But this is not the only way for these data to be produced. The experimenter may

[9] It means that Neyman and Pearson approach considers (only) two possible hypothesis *H* and H_a.

[10] In a Neyman-Pearson test you have to choose which of the two is your null-hypothesis and that choice may change which of the two is rejected (cf. Howson and Urbach 2006, 156).

have planned to flip the coin until he produced 3 heads, or until he becomes bored with the flipping. In this case, the outcome space will be different, even infinite or ambiguous, and the final result of the significance test may also be different. (cf. Loredo 1990, 109-110). It seems that the frequentist approach is more subjective than the Bayesian approach, because the identifications of outcome spaces, the choices of test statistics, the declarations of rejection regions, the choices of null-hypothesis among alternatives, the contradictory choices between sizes and powers etc., depend on thoughts or even whims of the experimenter. Frequentists thus failed to solve the problems that motivated their approach, they even exacerbated them.

The basic problem of frequentist analysis is that, in search of a rejection region, it evaluates a single hypotheses by taking into account data that *could have happened*. But what this *possible data* have to do with our problem? We have made our experiment, we have got *the real data* and we want to estimate hypotheses given *this real data*.

The result of the frequentis analysis is a behavioural attitude towards a single hypotheses, prompted by data that could have occurred but did not.

To be more specific, for Bayesians there is the probability of *H* being in an interval:

$$pr(R/N - 2/\sqrt{N} \leq H \leq R/N + 2/\sqrt{N}\ ) \approx 95\%$$

For frequentists there is no such probability. There is only the inductive behaviour according to which, when we prove that:

$$pr(H - 2/\sqrt{N} \leq H \leq R/N + 2/\sqrt{N}\ ) \approx 95\%$$

then if we behave so that we accept *R*/*N* as our estimate of *H*, we may expect to be correct in 95% of our repeated behaviours.

The simple Bayesian 95% probability that your hypothesis is true, is replaced by the convoluted frequentist 95% chance of being correct in your repeated "as if my hypothesis is being true" behaviours. Why on earth would anybody do that? Is there not a better answer to the frequentist critique that degrees of belief are subjective and therefore incoherent and (even if they have some sense) that we do not know whether they satisfy probability axioms. A lot of people thought there is.

For John M. Keynes a degree of rational belief is a degree of partial entailment. Sometimes a conclusion *follows* from premises, but more often it only *partially follows* from them. As Keynes used to say, a conclusion stands in a relation of probability with premises (cf. Keynes 1921, 52-3). The relation is logical, and probability is just an extension of classical "true or false" logic. But how do we asses this logical relation of probability and, more specifically, how do we establish the probability axioms from this logical point of view? Keynes thought we simply perceive them as true, with some kind of logical intuition. (cf. Keynes 1921, 52-3).

Harold Jeffreys held the same logical attitude towards probability. He was one of the earliest critics of the frequentist statistics, but he did more than criticize. In his Jeffreys of 1939, he solved many statistical problems completely inaccessible to frequentists. That should have been a clear indication that he was on the right track, even though his first hundred pages devoted to logical arguments for probability axioms were not very successful. His work was rejected on philosophical grounds, as was Keynes'.

The most famous critic was Frank Ramsey. His answer to Keynes' position (that there are logical relations of probability and that these can be perceived with some kind of logical intuition) was simple and final. He does not perceive the probability relations of Mr. Keynes and, moreover, he suspects that others do not perceive them either (cf. Ramsey 1926, 161-2).

I suppose Ramsey was referring to the probability axioms:

(1) $pr\colon \mathbf{P} \times \mathbf{P} \rightarrow [0,1]$

(i. e. probability is a real number from [0,1] assigned to an ordered pair of propositions[11] in $\mathbf{P} \times \mathbf{P}$, which measures how probable is the first proposition given the second),

(2) $\vDash A \Rightarrow pr(A|I) = 1$

(i.e. probability of a logically valid proposition is 1, whatever background information I),

(3) $I \vDash -(AB) \Rightarrow pr(A \vee B|I) = pr(A|I) + pr(B|I)$

(i.e. if A and B a contradictory given I, then the probability of their alternation, given I, is additive),

(4) $pr(AB|I) = \mathrm{pr}(A|B, I) \cdot pr(B|I)$

(i.e. the probability of conjunction is quasi-multiplicative).

I do not think that he or anybody else has problems with inferences of probability theory. For example, that from (1)-(3) it follows that $pr(A) = 1 - pr(A)$; or that from (1)-(3) and $A \vDash B$ it follows that $pr(A) \leq pr(B)$; or that from (1)-(4) follows Bayes' theorem; or that from (1)-(4) and $A \vDash B$ it follows that $pr(B|A, I) = 1$; or that from $pr(A|A{\rightarrow}B, I) \neq 1$ it follows that $pr(A|B, A{\rightarrow}B, I) > pr(A|A{\rightarrow}B, I)$[12]; etc.

Probability axioms are problematic, not probability inferences.

[11] Or statements, or sentences, here it is not important.

[12] Given $A{\rightarrow}B$; A does not follow from B, but A is more probable given B.

Instead of vague logical intuitions Ramsey (and de Finetti) offered a definition of probability and proved that it satisfies probability axioms (1)-(4). Namely, the probabilities were defined as betting quotients and it was proved that the betting quotients are coherent (i.e. do not allow unfair bets; which we define below) iff they satisfy (1)-(4). It was a great success and the logical theory was forgotten.

It is often declared that this is a very surprising result, and that it is by no means obvious that betting quotients, if coherent, should obey the probability axioms (c.f. Gillies 2000, 66). I think it is obvious, and I am offering an obvious proof bellow. Before that, let me present a more standard version of the proof that coherence implies axioms (1)-(4)[13].

Think of me as a bookie. If you are willing to pay me M' for prospect of getting M if A happens, then your net-gain $G(A)$ in this bet on A, is M–M' if A happens and –M' if it does not happen[14]. If we define the value of A as $V(A) = 1$ if A happens, and as $V(A) = 0$ if A does not happen, then

$$G(A) = MV(A) - M'.$$

If you are willing to pay me M' for prospect of getting M only if a condition C is fulfilled and A happens (i.e. the bet is cancelled if C is not fulfilled), then your net-gain $G(A|C)$ in this bet on A under condition C, is

$$G(A|C)) = V(C)(MV(A) - M')$$

(i.e. the bet is cancelled by $V(C) = 0$ and otherwise it is like before).

What I am offering you, i.e. M, is your possible brutto-gain or the value of the bet. What you are willing to pay for the bet, i.e. M', is your betting expectation. Your betting quotient, in this particular bet on A, is defined as

$$q(A) = M'/M.$$

In this definition it is presupposed that your expectation M' is proportional to the value of the bet M, i.e. that your betting quotient depends only on the proposition A you are betting on and not on M. Real bets are definitely not like that and this is the soft point of Ramsey-de Finetti's argument. But let's go further with the argument.

Since $M' = q(A)M$, your net-gains (c.f. above) can be reformulated as
$G(A) = M(V(A) - q(A))$
$G(A|C) = V(C)M(V(A) - q(A))$.

[13] The converse does not interest us here. In philosophical literature the proof of the converse is usually omitted, and if not omitted it is often incorrect (cf. Gillies 2000, 60-64; Hacking 2001, 165-168). Gilles "proves" that each of (1)-(4) taken separately, implies coherence. Of course, it is nonsense, because then each of them, taken separately, implies all of them, since coherence implies all of them. Hacking's proof is similar, although it is not clear whether he is claiming a proof or just an idea of it.

[14] Of course, a negative gain is a loss, as a negative loss is a gain.

Your betting quotients are said to be coherent (and your bets to be fair) if I can not choose my Ms so that I win whatever happens. Or, for that matter, that I can not choose them so that I lose whatever happens.[15] It means that the gain or the loss must depend on what happens. If it does not depend on what happens there is no gain and no loss, i.e. the gain is zero. More formally, *your betting quotients are coherent (and your bets are fair) if, and only if, G does not depend on V only if G = 0.*

Now, that we have defined coherence (fairness) we may prove that the probability axioms (1)-(4) follow from it.

Suppose that $q(A) \notin [0,1]$, i.e. $q(A) < 0$ or $q(A) > 1$. If $M > 0$, then $G(A) = M(V(A) - q(A)) > 0$ or $G(A) = M(V(A) - q(A)) < 0$ independently of the value $V(A)$. (If $M < 0$ then $G(A) < 0$ in the first case and $G(A) > 0$ in the second case.) Hence, $G(A) \neq 0$ independently of V, which is in contradiction with coherence. So, it is impossible that $q(A) \notin [0,1]$, i.e. $q(A) \in [0,1]$. This is our axiom (1).

If A is logically valid, i.e. $\vDash A$, then $V(A) = 1$ and

$$G(A) = M(V(A) - q(A)) = M(1- q(A))$$

does not depend on V. By coherence it must be zero, i.e. $M(1 - q(A)) = 0$, from which (for $M \neq 0$) it follows that $q(A) = 1$. This is our axiom (2).

If you bet on A with quotient $q(A)$ for brutto-gain M_1, and on B with $q(B)$ for M_2, and on $A \vee B$ with $q(A \vee B)$ for M; then your total net-gain is

$$G = M_1(V(A) - q(A)) + M_2(V(B) - q(B)) + M(V(A \vee B) - q(A \vee B)).$$

Now, if from your background information it follows that A and B are mutually contradictory, then $V(A \vee B) = V(A) + V(B)$. If furthermore, your bet is such that $M_1 = M_2 = - M \neq 0$ then, for this particular bet,

$$G = Mq(A) + Mq(B) - Mq(A \vee B).$$

This gain does not depend on V so, by coherence, it must be zero,

$$M(q(A) + q(B) - q(A \vee B)) = 0.$$

It follows that $q(A \vee B) = q(A) + q(B)$. This is our axiom (3).

If you bet on AB with quotient $q(AB)$ for brutto-gain M, and on B with $q(B)$ for M_1, and on A under condition B with quotient $q(A|B)$ for brutto-gain M_2; then your total net-gain is

$$G = M(V(AB) - q(AB)) + M_1(V(B) - q(B)) + V(B)M_2(V(A) - q(A|B)).$$

[15] Of course, changing the signs of my Ms turns one into another.

If your bet is such that $M_2 = - M \neq 0$ then, since $V(AB) = V(A)V(B)$, your net-gain is

$$G = - Mq(AB) + M_1V(B) - M_1q(B) + V(B)Mq(A|B).$$

If furthermore $M_1 = - Mq(A|B)$ then, fort this particular bet,

$$G = - Mq(AB) + Mq(A|B)q(B).$$

This gain does not depend on V so, by coherence, it must be zero,

$$M(- q(AB) + q(A|B)q(B)) = 0.$$

It follows that $q(AB) = q(A|B)q(B))$. This is our axiom (4).

This is a standard, maybe not extremely obvious proof. Now I present basically the same proof, which is trivial and completely obvious. Instead from coherence, I start from its simple consequence: for same bets you should have same expectations.[16] I define two bets as the same, if your brutto-gain in every possible situation is the same for both bets.

Example I: if A and B are mutually contradictory, then "to bet on $A \vee B$ for M" is the same as "to bet on A for M and to bet on B for M". Namely, there are only three possible situations $A\bar{B}$, $\bar{A}B$ and $\bar{A}\bar{B}$ (because AB is excluded) and in each of them your brutto-gain is the same for both bets (M if $\bar{A}B$ or $A\bar{B}$ and 0 if $\bar{A}\bar{B}$).

Example II: "to bet on AB for M" is the same as "to bet on B for M and then continue to bet on A for what you have got". Now, there are four possible situations, AB, $\bar{A}B$, $A\bar{B}$ and $\bar{A}\bar{B}$. In both bets you brutto-gains are the same in every of the four situations. They are, respectively: M, 0, 0, 0.

According to the example I, what you are willing to pay for bet on $A \vee B$ with brutto-gain M (if A and B are mutually contradictory), must be the same as what you are willing to pay for two bets, one on A for brutto-gain M and another on B for brutto-gain M. It means that

$$q(A \vee B)M = q(A)M + q(B)M,$$

and (for $M \neq 0$) it immediately follows that $q(A \vee B) = q(A) + q(B)$. This is our axiom (3).

According to the example II, what you are willing to pay for bet on AB for brutto-gain M, must be the same as what you one willing to pay for bet on B for brutto-gain M, which continues with the bet on A for what you have got. It means that

[16] It is a simple consequence of coherence. Namely, if you bet on A for M, with different expectations M_1 and M_2, i.e. with different quotients q_1 and q_2, then I may offer you M for one quotient and $-M$ for another. Your total net-gain in this compound bet is: $G = M(V(A) - q_1) - M(V(A) - q_2) = M(q_2 - q_1)$, which is independent of V and different from zero (because $q_1 \neq q_2$ and we can take $M \neq 0$). Hence, your quotients q_1 and q_2 are not coherent.

$$q(AB)M = q(A|B)(q(B)M),$$

and (for $M \neq 0$) it immediately follows that $q(AB) = q(A|B)q(B)$. This is our axiom (4).

I think the arguments for the axioms (1) and (2) were obvious. If your $q(A) > 1$ you obviously lose whatever happens, and if your $q(A) < 0$ you obviously win whatever happens. For valid proposition A, whatever happens, you obviously win if your $q(A) < 1$, (because valid A happens, whatever happens).

So, betting quotients quite obviously satisfy the probability axioms. There are no surprises about that. I would even suspect that these simple arguments for axioms (1)-(4) were well known from the beginnings of probability theory, because they are really extremely simple. Perhaps the reason they were not published (if they were not) is that the betting quotients were problematic, because they were not well defined.

And still today, they are not well defined. Presumption that *M'*, which you are willing to pay for a bet, is proportional to *M*, which is the brutto-gain you are hoping for, is completely unsubstantiated. Even Ramsey was aware of that when he unsuccessfully tried to overcome the problem by introducing "ultimate goods" bets, instead of money bets, cf. Ramsey 1926, 173-176.[17]

The subjective Bayesianism of Ramsey and de Finetti did not solve the problems of the logical (or objective) Bayesianism of Keynes and Jeffreys. But Cox in the 1940' (cf. Cox 1946 and Cox 1961) provided the missing foundation for logical Bayesianism, which is today known as Bayesian probability theory, or BPT for short.[18] The intuitive appeal of BPT[19], the huge amount of successful results and its rigorous mathematical foundation provided by Cox and others, make it the best theory of probable inference we have. Hence, it is quite strange that it is not even mentioned in the recent philosophy textbooks devoted to the probable inference (e.g. Gillies 2000, Hacking 2001 and Mellor 2005). It is mentioned in Bayesian textbooks, e.g. Howson and Urbach 2006 which explicitly declares it as the best approach ("which begs fewest questions of all"), but even then the Cox's mathematical foundation is omitted because "it requires fairly sophisticated mathematics".

Although mathematics is a bit sophisticated, I will present a variant of the crucial proof; especially because printed proofs are rare, quite often not completely correct and they almost always presuppose more assumptions then necessary (cf. Cox 1946, Cox 1961 and Jaynes 2003).

Cox's idea was to start from the notion of the *plausibility of a proposition A given a proposition I as known*, which is denoted by $A|I$, and from some properties that these plausibilities have to satisfy. I will use the following properties.

[17] Gillies proposes, following early de Finetti, that we should use money bets with appropriately selected stakes, with no real explanation, not to talk about a definition, of appropriateness, cf. Gillies 2000, 57.

[18] Jaynes call it "probability theory as the logic (of science)", cf. the title of Jaynes 2003.

[19] Just compare the Bayesian vs. frequentist analysis of the coin fairness problem above.

(P1) $\perp|I = o \leq A|I \leq I|I = j$

(i.e. plausibilities are real numbers between the minimum o, which is the plausibility of a logical contradiction and the maximum j, which is the plausibility of a logical truth).

(P2) $I \vDash -(AB) \Rightarrow A \vee B|I = AI(A|I, B|I)$

(i.e. if A and B are mutually contradictory given I, then the plausibility of their alternation "A or B", given I, is determined by the plausibility of A, given I, and the plausibility of B, given I; the determination function AI may depend on I).

(P3) Functions $\mathcal{A}_I$ are continuous and strictly increasing in both arguments.

(P4) $AB|I = KI\ (A|BI, B|I)$

(i.e. the plausibility of the conjunction "A and B" given I, is determined by the plausibility of B, given I, and the plausibility of A, given B and I; the determination function $\mathcal{K}_I$ may depend on I).

(P5) $AI|I = A|I.$

From these properties, using the logical rule of the replacement of equivalents (e.g. from $(A \vee B)C \equiv AC \vee BC$; it follows $(A \vee B)C|I = (AC \vee BC)|I$, from $I \equiv II$ it follows $A|I = A|II$ etc.) it is possible to prove that there exists a continuous and strictly increasing function $f(x)$ such that $f(o) = 0, f(j) = 1$ and that for every proposition I:

$$\mathcal{A}_I(x,y) = f(f^{-1}(x) + f^{-1}(y)), \qquad \mathcal{K}_I(x,y) = f((f^{-1}(x) \cdot f^{-1}(y)).$$

This is equivalent to:

$$f^{-1}\mathcal{A}_I(x,y) = f^{-1}(x) + f^{-1}(y), \qquad f^{-1}\,\mathcal{K}_I(x,y) = f^{-1}(x) \cdot f^{-1}(y).$$

Hence, if we define $pr(A|I) := f^{-1}(A|I)$ and substitute $A|I$ for x and $B|I$ for y, we get:

$$pr(A \vee B|I) = pr(A) + pr(B), \qquad pr(AB|I) = pr(A|BI) \cdot pr(B|I).$$

The conclusion is: if plausibility satisfies (P1)-(P5) then there is a measure of plausibility which satisfies our probability axioms (1)-(4). Namely, every continuous and strictly increasing function of plausibility $A|I$ could be a measure of plausibility, as any other. Out of all these possible measures we chose $pr(A|I)$, not because it is more "correct" but because it is more convenient, i.e. the quantities pr obey the simplest rules of combination: the normality condition (1), (2), the sum rule (3) and the product rule (4).

The situation is analogous to that in thermodynamics (cf. Jaynes 2003, 42), where out of all temperature scales (which are continuously increasing functions of each other) we choose Kelvin scale because it is more convenient, i.e. the laws of thermodynamics take the simplest form in this scale. Similarly, in mathematics, out of all angle scales we choose the radians as the most convenient; e.g. $d\sin x / dx = \cos x$ only if x is measured in radians.

Before I present the proof of this crucial result (usually called Cox's theorem) I should address one more problem. Why plausibilities should obey the properties (P1) – (P5)?

Desideratum (P1) is that degrees of plausibility are represented by real numbers (with the minimum which represent the plausibility of contradictions and the maximum which represent the plausibility of tautologies). I believe it is possible to prove that this desideratum follows from more elementary desiderata that (i) degrees of plausibility should be linearly ordered (i.e. that they are transitive, antireflexive and universally comparable), and that (ii) continuous, strictly increasing, commutative and associative operations (representing degrees of plausibility of conjunctions and alternations c.f. below) are definable on these degrees.[20]

In the moment it is just a conjecture, and I will not further discuss (P1).

As a first point about (P2), note that, given the knowledge of I, the process of deciding that $A\vee B$ is true, can be broken down into elementary decisions about A and B separately:

(i) Decide that A is true. $(A|I)$
(ii) Decide that B is true. $(B|I)$

In each step I indicate (in the brackets) the plausibility corresponding to that step. These two decisions completely determine our decision about $A\vee B$. More formally:

$$A\vee B|I = \mathcal{A}_I(A|I, B|I),$$

which is our (P2). Of course, if the plausibility in any of the two steps is increased then the combined plausibility of $A\vee B$ is increased, which is our (P3).

As for (P4), note that, given the knowledge of I, the process of deciding that AB is true can be broken into elementary decisions about A and B separately, in the following way:

(i) Decide that B is true. $(B|I)$
(ii) Having accepted B as true, decide that A is true. $(A|BI)$

Equivalently
(i') Decide that A is true. $(A|I)$
(ii') Having accepted A as true, decide that B is true. $(B|AI)$

Regarding the first procedure, in order for AB to be true it is necessary that B is true. So, $B|I$ is to be decided. Further, if B is true it is necessary that A is true. So, $A|BI$ is to be decided,

[20] The proof would be on adaptation of Hölder-Cartan proof that every linearly ordered group without minimum is embeddable in $\mathbb{R}$, and isomorphic to $\mathbb{R}$ if it is Dedekind continuous.

too. These two decisions completely determine our decision about AB. More formally:

$$AB|I = \mathcal{K}_I(B|I, A|BI),$$

which is our (P4)[21]. Of course, (P5) is self-evident.

If we define $x{:=}A|II$ and take into account that $j = I|I$, $I \equiv II$ and $AI \equiv A$, given I, then

$$\mathcal{K}_I(x,j) = \mathcal{K}_I(A|II, I|I) = AI|I = A|I = A|II = x.$$

Similarly,

$$\mathcal{K}_I(j,x) = \mathcal{K}_I(AI|AI, A|I) = AIA|I = AI|I = A|I = x.$$

In other words j is a neutral element for $\mathcal{K}_I$ (for every I).

It is as easy to prove that o is a neutral element for $\mathcal{A}_I$ (for every I):

$$\mathcal{A}_I(o,x) = \mathcal{A}_I(\bot|I, A|I) = (\bot\vee A)|I = A|I = x,$$

and similarly

$$\mathcal{A}_I(x,o) = \mathcal{A}_I(A|I, \bot|I) = (A\vee\bot)|I = A|I = x.$$

That $\mathcal{A}_I$ is associative, is proved in the following way:

$$\mathcal{A}_I(\mathcal{A}_I(x,y),z) = \mathcal{A}_I(\mathcal{A}_I(A|I,B|I),C|I) = \mathcal{A}_I((A\vee B|I,C|I) = ((A\vee B)\vee C)|I = (A\vee(B\vee C)|I$$
$$= \mathcal{A}_I(A|I, (B\vee C)|I) = \mathcal{A}_I(A|I, \mathcal{A}_I(B|I, C|I)) = \mathcal{A}_I(x, \mathcal{A}_I(y,z)).$$

It is even easier to prove that it is commutative, i.e. that:

$$\mathcal{A}_I(x,y) = \mathcal{A}_I(y,x).$$

Furthermore, $\mathcal{K}_I$(is distributive with respect to $\mathcal{A}_I$, i.e. for every I and C:

$$\mathcal{K}_I[\mathcal{A}_{CI}(x,y),z] = \mathcal{A}_I[\mathcal{K}_I(x,z), \mathcal{K}_I(y,z)].$$

Namely, $(A\vee B)C \equiv AC\vee BC$, hence

[21] In many discussions of uncertain reasoning (most prominently in AI discussions of fuzzy logics) it is quite common to suppose that $AB|I = \mathcal{K}(A|I, B|I)$, with various candidates for $\mathcal{K}$, although it is evident that no relation of this form is generally valid. (So, the discussions based on this assumption are completely futile.) For example, the plausibility of the next person being female and the plausibility of the next person being male could be about 50%, although the plausibility of the next person being male and female is zero. On the other hand the plausibility of the next person being older than 20 years and the plausibility of the next person being younger than 60 years could also be about 50%, although, in this case, the plausibility of the next person being older than 20 years and younger than 60 years should not be zero.

$(A \vee B)C|I = (AC \vee BC)|I.$

It follows that

$$\mathcal{K}_I((A \vee B)|CI, C|I) = \mathcal{A}_I(AC|I, BC|I),$$

which means that

$$\mathcal{K}_I[\mathcal{A}_{CI}(A|CI, B|CI), C|I] = \mathcal{A}_I[\mathcal{K}_I(A|CI, C|I), \mathcal{K}_I(B|CI, C|I)].$$

If we define $x := A|CI$, $y := B|CI$, and $z := C|I$, we finally have

$$\mathcal{K}_I[\mathcal{A}_{CI}(x,y),z] = \mathcal{A}_I[\mathcal{K}_I(x,z), \mathcal{K}_I(y,z)],$$

which was to be proved.

If we substitute $z = j$ in the above formula of distributivity we get:

$$\mathcal{K}_I[\mathcal{A}_{CI}(x,y),j] = \mathcal{A}_I[\mathcal{K}_I(x,j), \mathcal{K}_I(y,j)],$$

which simplifies to

$$\mathcal{A}_{CI}(x,y) = \mathcal{A}_I(x,y),$$

(because j is a neutral element of $\mathcal{K}_I$). We may repeat this while exchanging C and I and get

$$\mathcal{A}_C(x,y) = \mathcal{A}_{CI}(x,y) = \mathcal{A}_I(x,y).$$

The conclusion is that $\mathcal{A}$ does not depend on I. (That $\mathcal{K}$ does not depend on I, will follow from what follows.)

So far we have proved that:

$$\begin{array}{l} o \leq \mathcal{A}(x,y) \leq j \quad o \leq \mathcal{K}(x,y) \leq j \\ \mathcal{A}(x,o) = \mathcal{A}(o,x) = x \\ \mathcal{K}(x,j) = \mathcal{K}(j,x) = x \\ \mathcal{A}(\mathcal{A}(x,y),z) = \mathcal{A}(x,\mathcal{A}(y,z)) \\ \mathcal{K}[\mathcal{A}(x,y),z] = \mathcal{A}[\mathcal{K}(x,z), \mathcal{K}(y,z)] \end{array}$$

(where $\mathcal{K}$ could be any $\mathcal{K}_I$).

In what follows the binary operation $\mathcal{A}(x,y)$ is renamed $x \circ y$. This operation is defined on $[o,j]$, it is continuous, associative, commutative, strictly increasing in both arguments, and it has a neutral element o (in algebra usually called zero).

For any number u, such that $o < u < j$, we have

$$o < u < u \circ u < u \circ u \circ u < \ldots$$

(because ∘ is strictly increasing). Hence, if we define $\underline{1}\,u := u$, $\underline{2}\,u := u \circ u$, $\underline{3}\,u := u \circ u \circ u$, etc. We immediately see that $\underline{m}u < \underline{n}u$, whenever $m < n$.[22] Furthermore, if $u < v$ then $u \circ u < v \circ v$, $u \circ u \circ u < v \circ v \circ v$ etc. (because ∘ is strictly increasing). Hence, $\underline{m}u < \underline{m}v$, whenever $u<v$. In other words, the two valued function $\underline{m}u$ is continuous (because ∘ is continuous) and strictly increasing in both arguments ($m \in N$ and $u \in [o,j]$).

If we fix the first argument, i.e. m, we get the strictly increasing function $\underline{m}u$, of one argument u. Because of

$$j = \lim_{u \to j} u \leq \lim_{u \to j} \underline{m}u \leq j,$$

it follows that this function maps $[o,j]$ onto $[o,j]$. And it makes it in 1-1 fashion, because it is strictly increasing. Hence, for every $u \in [o,j]$ there is exactly one $v \in [o,j]$ such that $\underline{m}v = u$. We symbolize this v with

$$v := \frac{u}{\underline{m}}.$$[23]

Now we are ready to define our function f. For every $m/n \in [0,1]$

$$f\left(\frac{m}{n}\right) := \underline{m}\,\frac{j}{\underline{n}}\,.$$

Of course, we have to prove that for every km/kn

$$\underline{km}\,\frac{j}{\underline{kn}} = \underline{m}\,\frac{j}{\underline{n}}$$

It is easy if we note that $j/\underline{kn} = (j/\underline{n})/\underline{k}$, (this follows immediately from $\underline{kn}z = \underline{k}(\underline{n}z)$, which is obvious). Namely, $\underline{km}(j/\underline{kn})$ is equal to

$$\frac{j}{\underline{kn}} \circ \ldots \circ \frac{j}{\underline{kn}} \qquad (km \text{ times}),$$

Which is equal to

$$\left(\frac{j}{\underline{kn}} \circ \ldots\right) \circ \ldots \circ \left(\frac{j}{\underline{kn}} \circ \ldots\right) \qquad (m \text{ brackets; } k \text{ times in brackets}).$$

[22] Note that $\underline{n}u$ is not nu. By underlining n we stress the difference.

[23] Note that $u/\underline{n}$ is not u/n. By underlining n we stress the difference.

But then, each bracket is equal to $\underline{k}((j/\underline{n})/\underline{k})$, which is j/n. Hence, the whole value is equal to $\underline{m}(j/\underline{n})$, which was to be proved.

Now it is easy to prove some important properties of f. First of all, f is strictly increasing,

$$m_2 > m_1 \Rightarrow f\left(\frac{m_2}{n}\right) > f\left(\frac{m_1}{n}\right),$$

because ∘ is strictly increasing and $j/\underline{n} > o$:

$$f\left(\frac{m_2}{n}\right) = \left(\frac{j}{\underline{n}}\circ\ldots\circ\frac{j}{\underline{n}}\right)\circ\left(\frac{j}{\underline{n}}\circ\ldots\circ\frac{j}{\underline{n}}\right)\circ(o\circ\ldots\circ o) = f\left(\frac{m_1}{n}\right),$$

(in the first and the third bracket we have m_1 times $j/\underline{n}$; in the second and the forth bracket we have $(m_2 - m_1)$ times $j/\underline{n}$ and o). Furthermore,

$$f\left(\frac{m_1}{n}\right)\circ f\left(\frac{m_2}{n}\right) = \underline{m_1}\frac{j}{\underline{n}}\circ\underline{m_2}\frac{j}{\underline{n}} = (m_1 + m_2)\frac{j}{\underline{n}} = f\left(\frac{m_1 + m_2}{n}\right)$$

i.e. f is ∘-additive.

So, f is strictly increasing, ∘-additive function defined on rational numbers from [0,1], such that $f(0) = o$ and $f(1) = j$. There is the unique continuous ∘-additive extension of this function to the real numbers from [0,1] (remember, we presupposed that $\mathcal{A}$, which means ∘, is continuous). This extension, which we continue to denote f, is also ∘-additive:

$$f(x)\circ f(y) = f(x+y).$$

If we substitute $u = f(x)$ and $v = f(y)$ this is equivalent to

$$u\circ v = f(f^{-1}(u) + f^{-1}(v)).$$

Let us pause and state what we have proved so far.

There is a continuous and strictly increasing function f, defined on [0,1], such that f(0) = o, f(1) = j and
$$\mathcal{A}(u,v) = f(f^{-1}(u) + f^{-1}(v)).$$

If we substitute this result into distributive law

$$\mathcal{K}_1[\mathcal{A}(x,y),z] = \mathcal{A}[\mathcal{K}_1(x,z), \mathcal{K}_1(y,z)]$$

(which we proved above), we get:

$$\mathcal{K}_I[f(f^{-1}(x)+f^{-1}(y)),z] = [f^{-1}(\mathcal{K}_I(x,z)) + f^{-1}(\mathcal{K}_I(y,z))].$$

Further, if we denote $f^{-1}(x)$ by u, and $f^{-1}(y)$ by v, and apply f^{-1} to both sides of the above equations, then

$$f^{-1}[\mathcal{K}_I(f(u+v),z)] = f^{-1}[\mathcal{K}_I(f(u),z)] + f^{-1}[\mathcal{K}_I(f(v),z)].$$

We further simplify by defining $M(u,z) := f^{-1}[\mathcal{K}_I(f(u),z)]$, to get

$$M(u+v,z) = M(u,z) + M(v,z).$$

It means that M is additive in the first argument, from which it follows it is linear in the first argument (because it is continuous):

$$M(u,z) = k(z)u.$$

From the defining equation $M(u,z) = f^{-1}[\mathcal{K}_I(f(u),z)]$ it follows:

$$\mathcal{K}_I((f(u),z) = f(M(u,z)) \text{ i.e.}$$

$$\mathcal{K}_I(t,z) = f(M(f^{-1}(t),z)) = f(k(z)f^{-1}(t)).$$

Substituting j for t we get:

$$z = \mathcal{K}_I(j,z) = f(k(z)f^{-1}(j)) = f(k(z)),$$

from which it immediately follows that $k(z) = f^{-1}(z)$. Hence, (for every I),

$$\mathcal{K}_I(t,z) = f(f^{-1}(t) \cdot f^{-1}(z)).$$

Let us summarize what we have proved so far.

There is a continuous and strictly increasing function f,
defined on [0,1], such that f(0) = o, f(1) = j and

$$\mathcal{A}(u,v) = f(f^{-1}(u) + f^{-1}(v)).$$
$$\mathcal{K}(u,v) = f(f^{-1}(u) \cdot f^{-1}(v))$$

or equivalently

$$f^{-1}(\mathcal{A}(u,v)) = f^{-1}(u) + f^{-1}(v)$$
$$f^{-1}(\mathcal{K}(u,v)) = f^{-1}(u) \cdot f^{-1}(v).$$

If we substitute concrete plausibilities for u and v we get:

$$f^{-1}(\mathcal{A}(A|I, B|I)) = f^{-1}(A|I) + f^{-1}(B|I)$$
$$f^{-1}(\mathcal{K}(A|I, B|AI)) = f^{-1}(A|I) \cdot f^{-1}(B|AI).$$

Now we can define probability function *pr* by $pr(A|I) := f^{-1}(A|I)$ and finally get:

$$pr(A \vee B|I) = pr(A|I) + pr(B|I),$$

$$pr(AB|I) = pr(A|I) \cdot pr(B|AI)$$

(of course, we presupposed that $I \vDash -(AB)$).

At the end we should address Halpern counterexample to Cox's theorem, cf. Halpern 1999. The crucial point is that the counterexample presupposes there is only finitely many probability values. But it is trivially true that for every *m/n* there is a proposition with probability *m/n*, e.g. "from urn with *m* white balls and *n–m* non-white balls a white ball will be drawn". Hence, it is as relevant to probability as any statement about finite structures is to arithmetic. You may explore finite structures and finite probability spaces and these are important subjects, but they do not provide us with counterexamples to arithmetic or probability. (For example, in a finite field of residues modulo 7 there is only finitely many primes but this has nothing to do with Euclid's theorem on infinitude of primes in ordinary arithmetic.)

REFERENCES

Cox, R. T. 1946. Probability, Frequency and reasonable Expectation. *American Journal of Physics* 14: 1-13.

Cox, R. T. 1961. *The Algebra of probable Inference*. Baltimore: John Hopkins Press.

Gillies, D. 2000. *Philosophical Theories of Probability*. London: Routledge.

Hacking, I. 2001. *Probability and inductive Logic.* Cambridge: Cambridge Univ. Press.

Halpern, J. Y. 1999. Cox's theorem revisited. *Journal of artificial intelligence research* 11: 429-35.

Howson, C. & Urbach, P. 2006. *Scentific reasoning – The Bayesian Approach.* Chicago: Open Court.

Jaynes, E. T. 2003. *Probability Theory – The Logic of Science*. Cambridge: Cambridge Univ. Press.

Jeffreys, H. 1939. *Theory of Probability*. Oxford & The Clarendon Press, 1961.

Keynes, J. M. 1921. *A Treatire on Probability*. London: Macmillan, 1963.

Loredo, T. J. 1990. From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophisics. In *Maximum Entropy and Bayesian Methods*, ed. P. F. Fongere. Dordrecht: Kluwer Academic Publishers.

Mellor, D. H. 2005. *Probability – A Philosophical Introduction*. London: Routledge.

Ramsey, F. P. 1926. Truth and Probability. In *The Foundations of mathematics and other Logical Essays*. London: Routledge, 1931.

Sivia, D. S. 1996. *Data Analysis*. Oxford: Oxford Univ. Press.

Received: January 23, 2013
Accepted: April 21, 2014

University of Zagreb
Faculty of Mechanical Engineering and Naval Architecture
Ivana Lučića 5
10002 Zagreb, Croatia
zvonimir.sikic@gmail.com