JÜRGEN HUMBURG

THE BAYES RULE IS NOT SUFFICIENT TO JUSTIFY OR DESCRIBE INDUCTIVE REASONING

1.

I will state in my report a proposition concerning the system of Richter which I found 20 years ago. Furthermore I will present some generalizations of the original proposition, especially a generalization concerning the systems of Carnap and the systems of de Finetti and Savage. The generalization concerning the systems of de Finetti and Savage has only the form of a conjecture which is a part of a research program of mine.

The above-mentioned proposition says that induction in all these systems is only possible if we accept some empirical hypothesis; it is impossible to explain within these systems how we get such a hypothesis. In a general sense no induction in these systems is possible.

Finally I will outline how from my point of view induction works.

2. THE SYSTEM OF RICHTER (R)

(a) *Direct Theory*: Richter postulates the existence of an objective probability in nature. He claims that for each experimental description there exists an objective probability for the possible outcomes of this experimental arrangement; in the following, for the sake of brevity, we use the word "experiment" instead of "experimental description".

Remark: We must presuppose that such an experimental description for which we postulate an objective probability in nature, has exactly *one* realization. Of course we could also consider experimental *schemata* which permit repetitions, but we cannot assume that the different instances of such a schema have a priori the same objective probabilities, because that is already an empirical hypothesis which we can not make a priori; we postulate the existence of an objective probability only for a *single* experiment, not for schemata of experiments. – There is some confusion in the paper of Richter on this point.

(b) Richter states some simple axioms that the probability of a disjunction may be calculated from the constituents of the disjunction;

Erkenntnis **26** (1987) 379–390.

from these axioms he proves that this postulated objective probability can be measured in such a way that the addition principle holds.

(c) Richter introduces the notion of physical independence which, briefly described, means that between two experiments there is no physical interaction. With a few simple axioms he shows that, for two experiments which are physically independent, the corresponding probabilities will be multiplied to get the probability of the conjunction of the outcomes of these experiments.

The objective probability of an event E in an experiment H we designate by $p(E \mid H)$; if no confusion is possible we simply write $p(E)$.

(d) If we consider an experimental description H, and E is a possible outcome of H, we also may consider the experiment $H \wedge \hat{E}$, which is the performance of H with the additional information that E has occurred. With a few axioms Richter shows that for the probabilities under the conditional experiment $H \wedge \hat{E}$ we have:

$$p(A \mid H \wedge \hat{E}) = p(E \,.\, A \mid H)/p(E \mid H),$$

which is the usual formula for the conditional probability. If no confusion is possible we write simply $p(A \mid E)$ instead of $p(A \mid H \wedge \hat{E})$.

(e) Richter points out that, with these notions alone, no induction is possible, i.e., that it is impossible to describe by these notions alone how we will get information about the unknown postulated objective probability. Richter shows in particular that the confidence test does not work without further presuppositions; it only works if we accept some hypothesis, usually called an a priori hypothesis, as very credible.

(f) *Indirect Theory*: Richter concludes that we must have, in addition to the notion of objective probability, a notion of a degree of credibility. That means we need a function ϕ which gives the degree of belief in an hypothesis about objective probability. To work out this notion is the aim of his indirect theory.

First he points out with a few axioms that we can measure this degree of belief in such a way that ϕ is additive. Without further foundation he assumes that it is further σ-additive.

(g) I must now offer some details: Our universe of experience we describe by a chain K of successive experiments H^i; $K = (H^1, H^2, H^3, \ldots)$. Each experiment H^i has the possible outcomes $x^i_1, \ldots, x^i_{k_i}$; for reasons of mathematical simplicity we presuppose in the following $k_i = k$. A possible objective probability π on K is given

by the following countable vector:

$$\pi = (p_i, p^j_i, p^{j_1 j_2}_i, \ldots)$$

with

$$p_i = p(x^1_i),\ p^j_i = p(x^2_i \mid x^1_j),\ p^{j_1 j_2}_i = p(x^3_i \mid x^1_{j_1} \,.\, x^2_{j_2}), \ldots.$$

The space of all this vectors we designate by Π. The belief-function ϕ is a function on a σ-field $\mathscr{P}$ over Π, namely the following product space:

$$\mathscr{P} = \bigotimes_{s \geqq 0} \mathscr{S}^{j_1 \cdots j_s};$$

$\mathscr{S}^{j_1 \cdots j_s}$ is the Borel space over the following simplex:

$$S^{\cdots} := \Big\{(p^{\cdots}_i)\colon \sum_{1 \leqq i \leqq k} p^{\cdots}_i = 1\Big\}.$$

(h) Now Richter asks in which way we modify this belief-function ϕ to a function ϕ^* if we make an observation E in the chain K. By quite simple axioms he deduces the following formula: $\mathrm{d}\phi^* = \pi(E)\,\mathrm{d}\phi(\pi)$; to express that, in this formula, $\pi(E)$ is a function of π and E is fixed, we sometimes write $E\langle\pi\rangle := \pi(E)$.

Richter does not presuppose that ϕ is normalized; in particular he admits that it is not possible to normalize ϕ at the very beginning of our observations, i.e. that we have $\phi(\Pi) = \infty$. If it is possible to normalize ϕ^*, we get the following formula:

$$\mathrm{d}\phi^* = E\langle\pi\rangle\,\mathrm{d}\phi(\pi) \Big/ \int E\langle\pi\rangle\,\mathrm{d}\phi(\pi).$$

In this form, Richter's rule is analogous to the Bayes rule, with the only difference that we distinguish between two kinds of probability, the "subjective" ϕ and the objective $\pi(E)$.

In the following we will call Richter's rule simply Bayes' rule.

(i) We need a further notion: If we have a belief-function ϕ, Richter thinks that it is possible to do an estimation of the unknown objective probability. By simple axioms we get the following estimation function χ, which Richter calls *chance*:

$$\chi(E \mid \phi) := \int E\langle\pi\rangle\,\mathrm{d}\phi(\pi) \Big/ \int \mathrm{d}\phi(\pi).$$

(j) If we observe the event E, we change ϕ into ϕ^* and get the following new estimator: $\chi(A \mid \phi \,.\, E) := \int \pi(A \mid E)\, \mathrm{d}\phi^* / \int \mathrm{d}\phi^*$.

(k) The estimator χ obeys the laws of subjective probability, i.e., fullfils the addition and multiplication principle; in the sense of Carnap, who explicates his confirmation-function as an estimator of objective probability, the chance χ is the general form of a confirmation-function.

(l) The usual confidence test is now possible if we assume that the a priori hypothesis $\mathscr{H}$ has a degree of belief $\phi(\mathscr{H}) \geqq 1-\epsilon$. If α is the error probability of this confidence test, we have a certainty at least as great as $(1-\alpha)(1-\epsilon)$ that this confidence test will be successful.

3.

Now the question arises: From where do we get the belief-function ϕ? Richter gives the following explanation (E): (see (R), V §20, p. 311):

(E) Die Aufstellung der Glaubwürdigkeitsgrade geschieht intuitiv durch überschlägige Anwendung von F zur Änderung einer Gleichgewichtung, die einem mehr oder weniger gut vorstellbaren Zustande völligen Nichtwissens entspräche, unter Verwendung unserer gesamten, gerade im Bewußtsein befindlichen Erfahrung anstelle eines Versuchsergebnisses. (F is the BAYES rule).

In other words:

> The belief degrees we get by applying intuitively the BAYES rule to an *equidistribution*, which corresponds to a more or less imaginable state of ignorance by using as event E our whole observational material.

4.

I will now show that this explanation (E) is wrong and that it is impossible to get a belief function ϕ by the means of Richter. A worked-out presentation of the following proof you may find in my Diplomarbeit (H_0).

An equidistribution on $\mathscr{P} = \otimes \mathscr{S}^{j_1 \cdots j_s}$ seems to be the product measure ϕ_0 of the equidistributions of the simplexes $S^{\cdots}$. Letting B be an element of a finite segment of the product space $\mathscr{P}$, we then have:

$$\phi_0(B) = \int_B \cdots \int \mathrm{d}p_1 \ldots \mathrm{d}p_{k-1}\, \mathrm{d}p_1^1 \ldots \mathrm{d}p_{k-1}^1 \ldots\ldots \mathrm{d}p_{k-1}^{k \cdots k} / |S|^N.$$

N is the number of simplices $S^{\cdots}$ over which we integrate. Because $S^{\cdots}$ has the dimension $k-1$ we integrate only from the variable $p_1^{\cdots}$ to the variable $p_{k-1}^{\cdots}$. I think in this form we see that ϕ_0 is an equidistribution which may correspond to ignorance.

Let us now assume that x_1^1 occurs. Then we get ϕ_0^* according to the Bayes rule as follows:

$$\phi_0^*(B) = \int_B \cdots \int p_1\, \mathrm{d}p_1 \ldots \mathrm{d}p_{k-1}\, \mathrm{d}p_1^1 \ldots \mathrm{d}p_{k-1}^{k \cdots k} / |S|^N.$$

Having performed the experiment H^1, we are interested in the subsequent chain $K' = (H^{2\prime}, H^{3\prime}, \ldots)$, $H^{i\prime} := H^i \wedge \hat{x}_1^1$; we therefore project the measure ϕ_0^* to the field $\mathscr{P}' := \mathscr{S}^1 \otimes \mathscr{S}^{11} \otimes \mathscr{S}^{12} \otimes \ldots \otimes \mathscr{S}^{1 j_1 \cdots j_s} \otimes \ldots$; this gives the measure ϕ', which is characterized as follows: Let B' be an element of a finite segment of $\mathscr{P}'$; then we have:

$$\phi'(B') := \phi_0^*((S, B'))$$

$$= \int_S \int_B \cdots \int p_1\, \mathrm{d}p_1 \ldots \mathrm{d}p_{k-1}\, \mathrm{d}p_1^1 \ldots \mathrm{d}p_{k-1}^{k \cdots k} / |S|^{N'}.$$

According to the theorem of Fubini we get

$$\phi'(B') = \alpha \,.\, \int_{B'} \cdots \int \mathrm{d}p_1^1 \ldots \mathrm{d}p_{k-1}^{k \cdots k} / |S|^{N'}$$

$$\text{with } \alpha := \int_S p_1 \mathrm{d}p_1 \ldots \mathrm{d}p_{k-1}.$$

We see that ϕ' is again the equidistribution.

We seem to have the following result: In the system of Richter the process of induction is impossible. If we know nothing, we cannot learn anything.

5.

Without proof I will state that the chance χ_0 which originates from the belief function ϕ_0 is the Wittgenstein-function c_∞ of Carnap's λ-system. For this function χ_0 we have the chance one that the relative frequencies of each atomic outcome will converge against $1/k$, or formally: $\chi_0(h_n(x_i^{..}) \to 1/k) = 1$.

If we know nothing we cannot have the certainty one that the relative frequencies converge against $1/k$. Therefore the function ϕ_0 is not adequate the state of ignorance. Hence we have the following result:

THESIS. *If we know nothing we cannot say anything, in particular we cannot give a probability evaluation; there does not exist a probability distribution which is adequate to knowing nothing: Each probability distribution makes some empirical claims.*

6.

We may still doubt the above results, firstly because the equidistribution on $\mathscr{P}$ is not a canonical notion but depends on the scale we apply to $\mathscr{P}$, secondly because we can doubt whether the notion of equidistribution is adequate at all and whether with another ϕ, induction in the system of Richter is nevertheless possible. We will therefore investigate the following formulas, by which I hope to have comprehended quite generally the question of whether in the system of Richter induction is possible.

(a) Let ϕ_n^* be the normalized modified belief function ϕ after n steps of observation and consequently n steps of modification.

Induction is possible in the system of Richter if there exists a belief function ϕ with the property: Relative to each possible objective probability π there is objective probability one that the belief function ϕ_n^* will converge against a belief function ψ which assigns this objective probability π the measure one; formally:

$$(*) \qquad \exists \phi : \forall \pi : \pi(\{\omega : \phi_n^* \xrightarrow{t} \psi \wedge \psi(\{\pi\}) = 1\}) = 1;$$

$\xrightarrow{t}$ means convergence in some topology, for instance the vague or the weak topology; as demonstrated in section (9), it is impossible to

demand the convergence for each argument; that would be nonsense.

(b) Another, I think weaker form, of this principle we get if we consider the chances. Induction in the system of Richter is possible if there exists a chance $\chi(A \mid \phi \,.\, \{\omega_n\})$ – the chance after n observations – which converges relative to each possible objective probability π with probability one against π; formally:

$$(**) \qquad \exists \chi(A \mid \phi \,.\, \{\omega_n\}) : \forall \pi : \pi(\{\omega : \chi(\,.\mid \phi \,.\, \{\omega_n\}) \xrightarrow{c} \pi\}) = 1;$$

ω_n is the observed outcome after realization of $H^1, \ldots, H^n$. The exact significance of the convergence $\xrightarrow{c}$ we will let open for the moment. I think that (**) will follow from (*) because the function $\pi(A \mid \{\omega_n\})$ is bounded by one and continuous and therefore by integrating (*) we will get (**).

Now I will proof that (**) is wrong.

First we must now make the following trivial assumption about the convergence $\xrightarrow{c}$: Let the vector π consist only of components 0 and 1. Let the vector π' consist only of components $1/k$. Then it is impossible for the chance $\chi(\,.\mid \phi \,.\, \{\omega_n\})$ that we have both

$$\chi \xrightarrow{c} \pi \text{ and } \chi \xrightarrow{c} \pi'.$$

Now let us assume that (**) is valid. We construct to each outcome ω of K a vector π_ω consisting only of components 0 and 1, which assigns exactly this outcome ω the probability one. Then we must have for each ω: $\chi(\,.\mid \phi \,.\, \{\omega_n\}) \xrightarrow{c} \pi_\omega$. Let $\pi' = (1/k, 1/k, \ldots)$. Then there does not exist any ω for which $\chi(\,.\mid \phi \,.\, \{\omega_n\}) \xrightarrow{c} \pi'$ and hence (**) is wrong. Because (*) implies (**), (*) is wrong.

We may repeat the thesis of (5).

Each belief function excludes some possibilities of nature. There is no vacuus belief function. It is not possible to conceive a prior distribution ϕ which corresponds to knowing nothing, and to take this ϕ as the Archimedian point from which by the Bayes rule we can do induction and give a foundation of our knowledge. Such a prior ϕ does not exist.

Induction by the BAYES rule alone is impossible. If we know nothing we cannot say anything, in particular we cannot give a probability distribution.

We may state that relative to these results the postulated objective probability remains a metaphysical concept, which we cannot relate to observational material.

In Section 9 I will give an example for which induction in the sense of the above principles in the system of Richter is possible, if we assume some hypotheses.

I do not think that the above results are reasons to reject the notion of objective probability. I think they only show that induction with the concepts of the system of Richter alone is impossible as well as that the introduction of the notion of objective probability is difficult.

7.

It is easy to generalize this result of Section (6) to the system of Carnap, because Carnap has explicated his confirmation-function as an estimator of objective probability. We can therefore state that there does not exist a c-function from which, by the multiplication principle – the Bayes rule –, induction generally can be explained.

If we consider for instance the function $c_\lambda(h, e)$, which is symmetric in h for each information e, we never have convergence to the real objective probabilities if these real objective probabilities are not symmetric.

Carnap realizes this and says in his reply to Putnam (Sch, 987 pp) that we must admit that the c-function depends on the different considered hypotheses, and I agree.

We have seen that it is not possible to admit all hypotheses and to describe induction generally by constructing a prior measure function m which covers all possible hypotheses and by then using the Bayes rule.

I want to state explicitly that this result of course is not an argument against the value of Carnap's theory; it merely shows that each inductive probability-function is only adequate if we presuppose some empirical hypotheses.

8.

It is difficult to generalize the above results to purely subjective probability systems like the systems of de Finetti or Savage, because in these systems the notion of objective probability is suppressed and we can not argue in the above way.

At the moment a full proof for the following proposition is not available. I will therefore present it only as a conjecture.

Conjecture: Each subjective probability function p on the chain K has some empirical content, excludes some possibilities of nature and cannot be presupposed in a state of ignorance.

If we know nothing, we cannot say anything, in particular we cannot give a probability distribution.

To explain this conjecture, we may consider de Finetti's exchangeable probabilities. As proved for instance in (H), these functions imply probability one for the assertion that the limit of relative frequencies exists; that is an empirical assertion which we cannot make a priori.

I think that in this way it is possible to prove generally that each probability function on K must make some empirical assumption, and therefore no prior distribution exists on whose application the Bayes rule generally describes inductive reasoning.

9.

We will now give an example for which Richter's system in the sense of the principles of Section (6) works, namely that the chain $K = (H^1, H^2, \ldots)$ is the independent repetition of an experiment H. We confine ourselves to the case $k = 2$.

Then the vector π is characterized by one number θ, the possible objective probability of the outcomes $x_{\ddot{1}}$. As ϕ we may choose the equidistribution on the unit interval.

If r is the number of occurrences of $x_{\ddot{1}}$ in ω_n, we have

$$\pi_\theta(\{\omega_n\}) = \theta^r(1-\theta)^{n-r};\ \mathrm{d}\phi_n^* = \theta^r(1-\theta)^{n-r}\,\mathrm{d}\theta.$$

The measure ϕ_n^* has the following density:

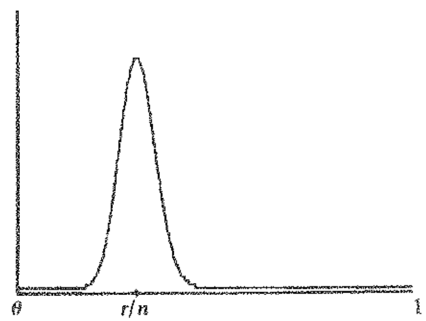

Fig. 1.

ϕ_n^* converges, informally speaking against a δ-function with increasing n.

Let ϕ_n^* be normalized; then we have, as is exactly proved in (H): If $r/n \to p$ then $\phi_n^*(\theta) \xrightarrow{d} D(\theta - p)$; D is the Dirichlet step function and $D(\theta - p)$ is the measure which exactly gives the point p the measure one. $\xrightarrow{d}$ symbolizes the distribution convergence. Further the strong law of large numbers holds, i.e., $\pi_\theta(r/n \to \theta) = 1$, and hence we have:

$$\pi_\theta(\{\omega : \phi_n^*(x) \xrightarrow{d} D(x - \theta)\}) = 1 \text{ for all } \theta.$$

Hence we see that under the assumption of independent repetition in Richter's system, induction in the sense of (6) is possible.

As mentioned in Section (6), it would be nonsense to demand in the above formula that $\phi_n^*(\{\theta\}) \to 1$; we must consider the distribution convergence $\xrightarrow{d}$.

10. HOW DOES INDUCTION WORK?

I will try to explain how from my point of view induction works. I use some notions of my book "Grundzüge zu einem neuen Aufbau der Wahrscheinlichkeitstheorie" (H); a short version in English you find in (H').

Let us consider for instance the experiment of throwing a die and let us presuppose that we know nearly nothing.

We may now assume for instance that this experiment has the following properties: By our fundamental sensory capacities, we always will realize that the successive throws of the die are *similar* processes; we may not perceive any differences in the successive throwing of the die; for instance we will not perceive that the die contains a clock changing the die.

Under this presupposition, I think we will be willing to form a concept, to make a hypothesis, namely the concept K of similarity, which means: We will think that throwing a 6 with the die today is as probable as to do it tomorrow; further we will think that to throw a 6 n times today will be as probable as to do it tomorrow.

From this concept K it follows, as shown in my book (H), that the limit of relative frequencies exists, especially that this concept is equivalent to the hypothesis of independent repetition.

Now we can make some probability prognosis concerning these experiments.

Let us now assume that we observe that often events occur which are very improbable. Then we may perhaps conceive a new hypothesis K', which we will compare with K in the following way:

We may give the two concepts a subjective probability, for instance to each 1/2, and then use the Bayes rule to get, by further observations, a decision between the two hypotheses. – And so we go on.

Of course I will not suggest that we must start with the concept of similarity, although it is especially simple, fundamental and frequent. – The forming of the concept K of similarity resembles the way in which Hume described inductive reasoning. – But of course a situation is possible where we start with another concept.

I think induction works in this way: We assume a hypothesis which seems reasonable relative to our elementary sensory perceptions and then go on in the described way.

Of course we will never get absolute certainty for an accepted hypothesis. We only can say that we have worked with an accepted hypothesis so successfully that we have never found another hypothesis which according to the Bayes rule we had to prefer to this accepted hypothesis.

I want to repeat my central result: It is not possible to consider at the very beginning of induction all thinkable hypotheses and to look for a prior on these hypotheses, corresponding to knowing nothing, from which then the Bayes rule gives us our knowledge.

The Bayes rule is not sufficient to justify or describe inductive reasoning.

ACKNOWLEDGEMENTS

For his contribution to this paper I thank Dr. E. Köhler. Furthermore I would like to thank the Deutsche Forschungsgemeinschaft for supporting this research.

REFERENCES

(C) Carnap, R. und W. Stegmüller: 1958, *Induktive Logik und Wahrscheinlichkeit*, Wien.

(C') Carnap, R. and C. R. Jeffrey: 1971, *Studies in Inductive Logic and Probability*, Vol. I, Berkley, Los Angeles, London.

(C″) Jeffrey, C. R.: 1980, *Studies in Inductive Logic and Probability*, Vol. II, Berkley, Los Angeles.

(E) Essler, W. K.: 1970, *Induktive Logik – Grundlagen und Voraussetzungen*, Freiburg, München.

(E′) Essler, W. K.: 1973, *Wissenschaftstheorie III – Wahrscheinlichkeit und Induktion*, Freiburg, München.

(F) Finetti, B. de: 1936, *La prévision: ses lois logiques, ses sources subjectives*, Vol. VII, 1, Institut H. Poincaré, Paris.

(F′) Finetti, B. de: 1938, 'Sur la condition d'équivalence partielle', *Colloque d'Octobre 1937 sur la Théorie des Probabilités*, Tome 6, Paris, pp. 5–18.

(F″) Finetti, B. de: 1981, *Wahrscheinlichkeitstheorie*, Wien, München.

(Hi) Hintikka, J.: 1966, A Two-Dimensional Continuum of Inductive Methods, Hintikka and Suppes (eds.), no. 7.

(H_0) Humburg, J.: 1965, Die Problematik des Induktiven Schließens bei Carnap und Richter, Diplomarbeit bei Prof. H. Richter an der Universität München.

(H) Humburg, J.: 1981, Grundzüge zu einem neuen Aufbau der Wahrscheinlichkeitstheorie, Frankfurt am Main.

(H′) Humburg, J.: 1986, 'Foundations of a New System of Probability Theory', *Topoi* **5**, 1.

(Hu) Hume, D.: *An Inquiry Concerning Human Understanding*, Section IV.

(R) Richter, H.: Zur Grundlegung der Wahrscheinlichkeitstheorie, (Math. Annalen: 125, 126, 128).

(S) Savage, L. J.: 1954, *The Foundations of Statistics*, New York.

(Sch) Schilpp, P. A.: 1964, 'The philosophy of Rudolf Carnap', *Library of Living Philosophers*, Vol. IX, London.

(St) Stegmüller, W.: 1971, Das Problem der Induktion: Humes Herausforderung und moderne Antworten, by Friedrich Vieweg, Braunschweig, Wissenschaftliche Buchgesellschaft, Darmstadt.

Schleissheimer Str. 173/I
D-8000 München 40
West Germany

Manuscript received 27 August 1986