Netherlands School of Economics

ECONOMETRIC INSTITUTE

Report 7217

STATISTICAL INFERENCE AND SUBJECTIVE PROBABILITIES

by J. Koerts and E. de Leede

Oktober 18, 1972

Preliminary and Confidential

STATISTICAL INFERENCE AND SUBJECTIVE PROBABILITIES

by J. Koerts[1] and E. de Leede[2]

Contents

SUMMARY

"Learning by experience" is a well-known part of the theory of subjective probabilities; the learning process is often derived from some prior distribution F(p) where p is a parameter of unknown value of a binomial process for instance. In this paper, the learning process is explicitly formulated and the corresponding prior distribution is derived from it. In this interpretation, subjective probabilities are part of an inference methodology, rather than a subjective evaluation of frequentistic probabilities. Implications are considered for a concept like the "non-informative prior"; the situation is considered in which the learning process seems to be in conflict with some objectively determined prior.

1. INTRODUCTION

The discussion between frequentists and subjectivists has left Dutch statisticians remarkably untouched. It seems that the subjective approach is thought to be a playground for psychologists. During the last few years, some econometricians have worked on it, undoubtedly because they are faced

[1] Netherlands School of Economics, Econometric Institute, Rotterdam.

[2] Interfaculty for Graduate Studies in Management, Rotterdam.

with situations where a limited number of observations is available, and where intuitive knowledge of some kind is present. One has to go back as far as 1957 if one wishes to find a contribution to the more fundamental discussion on these matters. In that year, Van Dantzig [2] published two articles in Statistica Neerlandica under the ominous title "Statistical Priesthood", in which he reviewed Savage's "The Foundations of Statistics" [10] and Fisher's "Statistical Methods and Scientific Inference" [6]. In particular, his criticism on the former is quite severe. Now, so many years later, the criticism does not always seem justified, and sometimes even unfair: Van Dantzig writes:

> "Combining the subjectivist's view with the statement in modern physics that radiation consists of probability waves, the reader would have to conclude that the inhabitants of Hiroshima and Nagasaki have been killed by waves of subjective degrees of expectation",

giving rise to the argument that the example is not very subtle (due to the intensity of the discussion?) and, worse, that it is unfair towards the theory of Savage. The personalistic view, as defended by Savage, does not deny the existence of stochastic phenomena in nature, but it does claim that our knowledge about such phenomena is essentially imperfect, so that a personalistic interpretation of data is (essentially) always necessary. There are no "waves of subjective degrees of expectation" that could have killed the people of Hiroshima and Nagasaki, but our knowledge about the phenomenon of radiation is (essentially) subjective.

The quotation is interesting, because it demonstrates a fundamental difference between subjective and objective probabilities: objective probabilities are related to natural phenomena, subjective probabilities are related to some person's knowledge about phenomena. This implies that the subjective probability approach is incomparable with the frequency interpretation of probabilities. If a comparison is possible, then it should be made between the subjectivistic approach and the inference methodology used by frequentists. In this light, it seems to be possible that the concept of stabilizing relative frequencies can be used together with the concept of subjective probabilities. The two concepts are complementary, rather than contradictory. In this paper, we shall

concentrate on the kind of problems, where complementarity might be expected. We are interested in sequences of observations, in which the n^{th} outcome is independent of any other outcome (although our opinion about the n^{th} outcome is influenced by our knowledge about other outcomes). This situation is typical for inference problems. The actual comparison between the subjectivistic approach and the inference methodology of the frequentists will be reserved for a second paper; here, we shall try to construct a subjective probability theory that can be used for inference problems. We shall restrict ourselves to trials with only two outcomes, notated as H and T (sometimes denoting heads and tails; the same notation is also used for other experiments).

2. THE SUBJECTIVISTIC APPROACH

Although subjective probabilities have been defined in very sophisticated ways (see for instance Savage [10]), it always boils down to the analysis of betting situations [7]. It is not necessary that bets form an actual part of the decision problem for which subjective probabilities must be evaluated; it is sometimes sufficient that a person evaluates subjective degrees of belief by means of hypothetical bets. In the latter case, the following assumption is made:

ASSUMPTION 1: If a decision problem under uncertainty is extended with hypothetical bets on the states of nature that might prevail, then the subjective probabilities defined on the basis of these bets will be valid for the original decision problem as well.

This assumption says that hypothetical bets are a good yardstick for the determination of subjective probabilities. A bet is considered fair if one and the same person is willing to lay a wager on both sides of the bet. Subjective probabilities can now be defined by means of fair bets in the following way:

If a person is willing to bet on an outcome A with stakes p : q, and if he is willing to bet on the complementary outcome A^c with stakes q : p, then the subjective probability of the outcome A can be defined as the ratio p/(p + q).

It may be noted that in this definition the utility of money is assumed to be linear. The utility concept can be avoided when defining subjective probabilities, but here we take the simplest possible definition, because it is not crucial to the later part of this paper.

The notion of fair bets is not sufficient for a reasonable definition of subjective probabilities. A person betting with stakes p : q on the outcome A, might possibly think that A ∪ B is worth a fair bet with stakes p' : q with p' < p. This inconsistency is impossible if the following assumption is made:

ASSUMPTION 2: Fair bets are acceptable only, if no combination of such bets can result in a certain loss.

The well-known rules of calculation for probabilities can be derived from the definition and assumption 2. It can be seen that $0 \leq P(A) \leq 1$ for any A; for the certain event, we find $P(S) = P(A) + P(A^c) = 1$. Finally, if $A \cap B = \emptyset$, then $P(A) + P(B) = P(A \cup B)$; indeed, suppose that $P(A \cup B) < P(A) + P(B)$ holds good. Consider then the three following bets: a bet on A with stakes P(A) : 1 - P(A); a bet on B with stakes P(B) : 1 - P(B); a bet on $(A \cup B)^c$ with stakes 1 - P(A ∪ B) : P(A ∪ B). If A prevails, these three bets will give a net result $(1 - P(A)) - P(B) - (1 - P(A \cup B)) < 0$. If B prevails, the net result will be $-P(A) + (1 - P(B)) - (1 - P(A \cup B)) < 0$. If neither A nor B prevail, the net result will be $-P(A) - P(B) + P(A \cup B) < 0$. The three bets together result in a certain loss. In the same way, if $P(A \cup B) > P(A)+P(B)$ holds good the complementary bets will result in a certain loss.

Next, consider two experiments, each of which results in either H or T: two coin tosses, for instance, where H stands for the outcome heads and T stands for the outcome tails. Probabilities could be assessed for the four points in the two-dimensional Cartesian product $\{(H_1, T_1) \times (H_2, T_2)\}$. Marginal probabilities can be derived from these probabilities by means of

$$P(H_2) = P(H_2T_1) + P(H_2H_1), \text{ etc}$$

Conditional probabilities are defined by

$$P(H_2/H_1) = \frac{P(H_2H_1)}{P(H_1)} \text{ , etc.}$$

Formally, this does not restrict the probability assessor in his original assessment of probability values in the two-dimensional space. It is quite possible, however, that direct assessment of probabilities in the second experiment alone, would result in values that differ from the marginal probabilities as derived above. Such behaviour is excluded in the theory of subjective probabilities, because of the following assumption:

ASSUMPTION 3: Given a sequence of n experiments with probabilities assessed to the n-tuples of possible outcomes, then any sequence of (n - k) experiments should be considered as a subsequence, i.e. as imbedded in the larger system, and probabilities concerning such a subsequence are marginal probabilities as derived from the n-dimensional Cartesian product.

This assumption guarantees consistency when sequence and subsequence are considered at the same moment.

Conditional probabilities can be interpreted as betting quotients for conditional bets, i.e. bets that are valid if a particular event occurs, and not valid (called off) if the event does not occur. The definition of conditional probabilities asserts that no combination of unconditional bettings can be made such that the total result is a certain loss. (Suppose, as a counter example, that $P(H_2H_1) = P(H_2T_1) = P(T_2H_1) = P(T_2T_1) = \frac{1}{4}$, and that $P(H_2/H_1)$ is equal to $4/5$ in the interpretation of a betting quotient in a conditional bet. Then a direct bet on the outcome T_2H_1 would be acceptable if the net gain for that outcome is 24 and the net loss on the three other outcomes is -8. Equally acceptable would be the conditional bet that gives a gain of +7 if H_2H_1 occurs, a loss of -28 if T_2H_1 occurs, and 0 if T_1 occurs (called off). Evidently, the combination of these two bets results in a certain loss. Conditional bets are defined at the moment that no observations have yet been made. The probability assessor may possibly change his mind after the first observation has been made. In that new situation, he could

accept his loss and reconsider his judgements. The following assumption excludes this behaviour.

ASSUMPTION 4: The probability P(A/B) of an event A after the hypothetical observation of B is equal to the probability that is given to A after the actual observation of B.

Essentially, the construction of probabilities after the observation of n outcomes (n = 1, 2, ...) is a dynamic process. The overall probability model is static. According to assumption 4, the static model must be constructed such that it contains the dynamic elements.

In the subjectivistic theory the notion of independent trials of an experiment is also introduced, but in an entirely different way as compared with the frequency theory. The idea of physically independent trials is reflected in the following assumption:

ASSUMPTION 5: If ordering numbers are assigned to all trials, then the ordering numbers can be interchanged within any probability statement.

In other words: if the outcomes can be described by the set of random variables $x_1, \ldots, x_n$ then any permutation of these variables has the same n-dimensional probability distribution as the original ordering $(x_1, x_2, \ldots, x_n)$. If this assumption is made then the random variables are called exchangeable or symmetrically dependent.

This assumption implies that we, in our personalistic beliefs, do not wish to learn from the ordering of the trials. In other words: physical independence is assumed for the sequence under consideration. Therefore, although the person does not know beforehand the results of future observations he maintains the hypothesis that it will not be a systematic sequence.

Although it is assumed that the n^{th} outcome does not influence any other future outcome, this does not imply that our knowledge about the n^{th} observation would not influence our subjective probability about other future outcomes. This fact is expressed by the expression symmetrically dependent. Hence the well-known multiplication rule $P[H_2 \cap H_1] = P[H_2] \cdot P[H_1]$, which holds good for independent collectives in the frequency theory, does not hold good if p denotes our subjective degree of belief.

The following important theorem about sequences of symmetrically dependent variables is due to de Finetti:

THEOREM: Given any infinite sequence of exchangeable variables x_n that can have values 0 or 1 only, there is a corresponding probability distribution F(p) on some parameter p in the interval [0, 1] such that the (subjective) probability on k successes in n experiments can be written as

$$P\ (k \text{ successes in } n \text{ experiments}) = \binom{n}{k} \int_0^1 p^k (1 - p)^{n-k} dF(p).$$

The distribution F(p) is unique almost everywhere. For the proof of this theorem, the reader is referred to [5] or [4].

This theorem gives us the opportunity of introducing the concept of a prior distribution in the subjectivistic theory. We shall first discuss two interpretations.

In the first interpretation, subjective and frequentistic probabilities are mixed. Suppose that a bag contains a proportion p of black balls and a proportion 1 - p of white balls. Experiments can be carried out by taking a ball from the bag, looking at the colour and returning it. Suppose now that one takes a quick look at the contents of the bag. Then one has some intuitive ideas about the true p-value. One could translate these ideas into a subjective probability distribution f(p) on [0, 1]. By taking out balls, in the aforementioned manner one could apply Bayes' theorem in order to obtain the posterior distribution of p:

$$f(p/\tfrac{k}{n}) = \frac{p^k (1 - p)^{n-k} f(p)}{\int_0^1 p^k (1 - p)^{n-k} f(p) dp}$$

in which $\frac{k}{n}$ in the left member denotes the observation of k successes in n experiments. It can be shown that for n going to infinity and for the observed relative frequency $r = \frac{k}{n}$ fixed, the variance of the posterior distribution will tend to 0, whereas the mean tends to r. For this property it is required there be at least one point in the neighbourhood of r, where f(p) will not vanish, and that f(p) is bounded (see, for instance, von Mises [8], who also gives some weaker requirements). In other words, if there is something like a true frequentistic probability r, we know that the subjective posterior probability will become equal to it if a sufficient number of experiments can be carried out.

The mixing of subjective probabilities and "true" frequentistic probabilities in this first interpretation does not seem to be very fortunate. For those subjectivists, who deny the sense of the frequentistic probability concept, it is inadmissible to introduce such a concept through the backdoor. It is not necessary, however, to interpret the parameter p as a frequentistic probability. If subjective probabilities are assigned to sequences of n outcomes (n = 1, 2, ...) such that (a) the probabilities on the n-dimensional space may be considered as marginal probabilities on a n-dimensional subspace of the (n + 1)-dimensional space; (b) the sequence is symmetric; then, the measure F(p) may be considered as a mathematical tool for which no interpretation is necessary. Formally, the frequentistic interpretation of F(p) can be avoided. It is worthy of note that there is still a connection, albeit in a difference sense, between frequencies and the distribution F(p). If r_n denotes the relative frequency of successes in n experiments, and if the subjective probability that is assigned to the event $r_n = \frac{k}{n}$ is written as

$$P[r_n = \frac{k}{n}] = \binom{n}{k} \int_0^1 p^k (1 - p)^{n-k} dF(p)$$

then the distribution function $F_n(x) = P(r_n \leq x)$ will tend to the prior distribution F(x) for n tending to infinity. Indeed, we have

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} \sum_{k=0}^{[nx]} \binom{n}{k} \int_0^1 p^k (1-p)^{n-k} dF(p) =$$

$$= \lim_{n\to\infty} \int_0^1 \sum_{k=0}^{[nx]} \binom{n}{k} p^k (1-p)^{n-k} dF(p) =$$

$$= \int_0^1 \lim_{n\to\infty} \sum_{k=0}^{[nx]} \binom{n}{k} p^k (1-p)^{n-k} dF(p)$$

As is well-known, the integrand will vanish for large n if $x < p$; the integral will equal unity for large n if $x \geq p$. So we get

$$\lim_{n\to\infty} F_n(x) = \int_0^x dF(p) = F(x)$$

The interpretation of F(p) as a limiting distribution of (subjective) probability distributions on the relative frequencies r_n will be called here the second interpretation of subjective prior distributions. For this second interpretation, it is not necessary that something like a "true" (frequentistic) probability should exist. Although the two interpretations are rather close, one should take care to distinguish between them; issues like the existence of "true" frequentistic probabilities have given rise to so much discussion that one may not neglect the subtle distinction between subjective probabilities on the unknown frequentistic probability value p (interpretation 1) and subjective probabilities on relative frequencies r_n for arbitrarily large values of n (interpretation 2). We shall treat the logical interpretation of probabilities before discussing the two interpretations. It will appear that this logical interpretation if formally a special case of the subjective approach in its second interpretation.

3. THE LOGICAL APPROACH

For the logical approach, we shall follow the train of thought of Carnap (see [1]). The theory is called "logical", because it describes a formal procedure for induction. The procedure itself is independent of the problems to which it is applied. There is no room for subjective "beliefs", as in the subjectivistic approach. Everyone could - if they wished - apply the methodology, and the results should then be the same

for all concerned. For that reason, Carnap uses the terminology "degree of confirmation" rather than "degree of belief". For comparison's sake, we shall often use the terminology that is known from the theory of statistics. It must be remarked, however, that Carnap claims a more wide-spread field of application for his logic. It should be a starting point for a more general theory of induction.

Carnap does not deny the possible existence of frequentistic probabilities. He calls it "probability$_2$" in order to distinguish it from betting quotients (or, if necessary, estimates of probability$_2$), to which he refers as probability$_1$. The theory of Carnap concentrates upon the structure of the latter kind of probabilities. We shall consider in brief the most important axioms underlying this structure.

The degree of confirmation that some hypothesis h is true, given the experience e, is denoted by c(h, e). The theory of Carnap takes into consideration sequences of experiments with only two outcomes.[1] We shall use the symbols H_i and T_i for the two possible outcomes of the i^{th} experiment, referring to coin tossing experiments. Both the hypothesis h and the experience e (the latter containing <u>all</u> available experience) must be given in the symbols H, T and i. If no experience is available, we shall write m(h) instead of c(h, e). It is assumed that the confirmation function is symmetric as to the outcomes H and T, i.e. in c(h, e) the symbols H and T may be interchanged without changing the value of the degree of confirmation. This implies m(H) = m(T); this assumption is known in other contexts as the principle of insufficient reasoning. The confirmation function should be insensitive to the sequence number of the experiments, i.e. in c(h, e) these ordering numbers may be interchanged without affecting the value of the degrees of confirmation. The third assumption is, that the confirmation function behaves like a probability measure: c(h, e) will always be between 0 and 1;
$c(h_1$, or h_2, e) = $c(h_1$, e) + $c(h_2$, e) if the conjunction $(h_1.h_2)$ is untrue; c(h or not -h, e) = 1; and, last but not least, the property of conditional probabilities is fulfilled, in other words,
c(h.j, e) = c(h, j.e).c(j, e).

Under these assumptions, the theory of Carnap is formally embedded in the subjective approach, where symmetric sequences of events have been defined. It is a special case, because symmetry is assumed to exist between the outcomes H and T; the latter kind of symmetry could be introduced on subjective grounds as well, and in that case the confirmation

---

[1] Strictly speaking, Carnap's theory is engaged with independent outcomes and their negatives, like for instance female/not female and blue eyes/not blue eyes. Kemeny and Carnap make a further extension, including such situations as the throwing of dice [1b].

function of Carnap seems to be identical with the specific subjective probability measure. The interesting point of Carnap's theory, however, is the construction of the confirmation function. According to the subjectivists, anybody is free in the construction of the probability measure, as long as the consistency requirements are fulfilled. According to Carnap, it should be possible to define a generally accepted methodology for induction.

Suppose that no experience is available. In that case the principle of insufficient reasoning results in $m(H) = m(T) = \frac{1}{2}$. Now suppose that experience e is available in terms of H, T and i only. The ordering numbers i are irrelevant, and only the observed relative frequency $\frac{k}{n}$ is of importance. We shall denote this by $\frac{k}{n}$, keeping in mind that n is the number of observations made (so $\frac{1}{2}$ is different from 2/4). Now Carnap assumes that

$$\frac{k}{n} \leq c(H, \frac{k}{n}) \leq \frac{1}{2} \quad \text{or} \quad \frac{1}{2} \leq c(H, \frac{k}{n}) \leq \frac{k}{n}$$

for $\frac{k}{n} \leq \frac{1}{2}$ and $\frac{k}{n} \geq \frac{1}{2}$ respectively.

In other words, $c(H, \frac{k}{n})$ is a convex combination of $m(H)$, having the value $\frac{1}{2}$, and the observed relative frequency $\frac{k}{n}$. We may write

$$c(H, \frac{k}{n}) = A.\frac{k}{n} + (1 - A).\frac{1}{2} = (1 - A).(\frac{A}{1 - A} . \frac{k}{n} + \frac{1}{2}) \quad (0 \leq A \leq 1)$$

Now the relative weight that is given to the observed relative frequency should increase with the number of observations made. Supposing that $\frac{A}{1 - A} = \frac{n}{\lambda}$ with λ a constant, we find the formula

$$c(H, \frac{k}{n}) = \frac{k + \lambda.\frac{1}{2}}{n + \lambda} \quad \text{with} \quad 0 \leq \lambda \leq \infty$$

For $\lambda = 0$, the degree of confirmation is determined by the observed relative frequency only. For $\lambda = \infty$ we see that the degree of confirmation is not influenced by the observed frequency; the value is $\frac{1}{2}$ and we do not wish to change our opinion in the light of any evidence. According to De Finetti's theorem, there is some prior distribution F(p) such that the degrees of confirmation are generated by a binomial process with probability p, conditional to the prior distribution F(p). For

constant values of λ, this prior distribution is a beta distribution with parameters $(\frac{\lambda}{2}, \frac{\lambda}{2})$. This can be verified by calculating the degree of confirmation m(k times H and n - k times T) directly in accordance with Carnap's theory and by calculating the binomial probability of k successes in n experiments from a prior that has a beta distribution with the parameters $(\frac{\lambda}{2}, \frac{\lambda}{2})$. Direct calculation gives for some specific sequence of outcomes (for instance the first n - k times T, the next k times H):

$$m(H_n.H_{n-1}. \ldots .H_{n-k+1}.T_{n-k}. \ldots .T_1) =$$

$$c(H_n, H_{n-1}. \ldots .T_1).m(H_{n-1}. \ldots .T_1) = \ldots =$$

$$c(H, \frac{k-1}{n-1}).c(H, \frac{k-2}{n-2}). \ldots .c(H, \frac{0}{n-k}).m(T_{n-k}. \ldots .T_1) =$$

$$\frac{k-1+\lambda.\frac{1}{2}}{n-1+\lambda} \cdot \frac{k-2+\lambda.\frac{1}{2}}{n-2+\lambda} \cdot \ldots \cdot \frac{0+\lambda.\frac{1}{2}}{n-k+\lambda} \cdot m(T_{n-k}. \ldots .T_1) =$$

Because $m(T_{n-k}. \ldots .T_1) = m(H_{n-k}. \ldots .H_1)$ which can be evaluated in the same way, we find as result

$$m(H_n . \ldots .H_{n-k+1}.T_{n-k}. \ldots .T_1) =$$

$$\frac{k-1+\lambda.\frac{1}{2}}{n-1+\lambda}.\frac{k-2+\lambda.\frac{1}{2}}{n-2+\lambda}. \ldots .\frac{0+\lambda.\frac{1}{2}}{n-k+\lambda}.\frac{n-k-1+\lambda.\frac{1}{2}}{n-k-1+\lambda}.\frac{n-k-2+\lambda.\frac{1}{2}}{n-k-2+\lambda}.$$

$$. \ldots . \frac{0+\lambda.\frac{1}{2}}{0+\lambda}$$

On the other hand, we can calculate

$$\frac{\int_0^1 p^k(1-p)^{n-k}p^{\frac{\lambda}{2}-1}(1-p)^{\frac{\lambda}{2}-1}\,dp}{\int_0^1 p^{\frac{\lambda}{2}-1}(1-p)^{\frac{\lambda}{2}-1}\,dp} = \frac{\Gamma(\frac{\lambda}{2}+k).\Gamma(\frac{\lambda}{2}+n-k)}{\Gamma(\lambda+n)} \cdot \frac{\Gamma(\lambda)}{\{\Gamma(\frac{\lambda}{2})\}^2}$$

and from the recursion formula $\Gamma(x) = (x-1).\Gamma(x-1)$ this appears to be equal to the directly calculated result.

Mathematically speaking, this result is not new. Beta-distributions are used as so-called conjugate priors, conjugate to a binomial process, and the results are explored quite thoroughly by subjectivists. The interesting part is the interpretation that is given by Carnap to the confirmation function. The value of λ denotes our willingness to learn from experience, and has nothing to do with subjective feelings. For $\lambda = 0$, our willingness to learn from experience is maximum; the corresponding prior distribution is a two-point distribution with half of its mass on the point $p = 1$. For $0 < \lambda < 2$, the density of the prior distribution is convex. For $\lambda = 2$, the prior distribution is rectangular, and for $\lambda > 2$, the density of the prior distribution is concave. According to Carnap, any choice of λ is permitted. The choice between a convex and a concave prior distribution is not dependent upon intuitive knowledge about the "true" value of p, but depends only upon our open-mindedness to experience obtained as observed relative frequencies.

We are now in the position where four different interpretations of the prior distribution are possible. These four interpretations are:

(a) the prior distribution reflects exact knowledge about a "true" distribution of p (an objective interpretation);

(b) the prior distribution reflects degrees of belief that a person has about the "true" value of p (mixed approach);

(c) the prior distribution is a mathematical idealization of personal degrees of belief given to relative frequencies in long sequences of events (subjective interpretation);

(d) the prior distribution is derived from some explicit learning process (constructive interpretation).

In Section 4 we shall consider the implications of the choice of prior distributions that are not beta-distributed for the learning process. Section 5 will be devoted to the special case of Carnap's confirmation function in which $\lambda = 0$ is chosen.

In Section 6 we shall compare the above mentioned interpretations, and arrive at some final conclusions.

4. CONFIRMATION FUNCTIONS WITH VARIABLE λ

Suppose that, on personalistic grounds, a prior distribution is chosen the density (or the mass function) of which is symmetric around $p = \frac{1}{2}$, whereas the prior is not beta-distributed. We may always write

$$P(H/\frac{k}{n}) = \frac{\int_0^1 p^{k+1}(1-p)^{n-k}dF}{\int_0^1 p^{k}(1-p)^{n-k}dF} = \frac{k + \lambda(k, n).\frac{1}{2}}{n + \lambda(k, n)}$$

from which $\lambda(k, n)$ may be solved as a function of k and n. The prior distribution defines a learning function that is similar to the confirmation function of Carnap, the only difference being that λ is no longer a constant. We may ask ourselves to what extent this learning function obeys the underlying ideas of Carnap's learning process. The symmetry between H and T is maintained in the prior distribution and will therefore present no problems. The physical independence of experiments is part of the probability model used and is found in the learning model as well. Difficulties are met only in the gradual effect of learning, as presupposed by Carnap. This gradual learning effect can be made explicit in the following four requirements:

(1) As Carnap pointed out, the requirements $\frac{k}{n} \leqq c(H, \frac{k}{n}) \leqq \frac{1}{2}$ or, $\frac{1}{2} \leqq c(H, \frac{k}{n}) \leqq \frac{k}{n}$ should be fulfilled. This boiled down to the requirement $0 \leqq \lambda \leqq \infty$.

Not mentioned by Carnap, but implicitly fulfilled in Carnap's confirmation function with λ a positive constant, are:

(2a) If $\frac{i}{m} \leqq \frac{k}{n} \leqq \frac{1}{2}$ with $m \geqq n$, then $c(H, \frac{i}{m}) \leqq c(H, \frac{k}{n})$;

(2b) If $\frac{i}{m} \geqq \frac{k}{n} \geqq \frac{1}{2}$ with $m \geqq n$, then $c(H, \frac{i}{m}) \geqq c(H, \frac{k}{n})$.

It can easily be verified that the requirements (2a) and (2b) are fulfilled in the case where λ is constant. Taking the first one, we have to prove that $\frac{i + \lambda.\frac{1}{2}}{m + \lambda} < \frac{k + \lambda.\frac{1}{2}}{n + \lambda}$, and this can be rewritten as the inequality $\frac{k}{n} - \frac{i}{m} + [(\frac{1}{2} - \frac{i}{m})\frac{\lambda}{n} - (\frac{1}{2} - \frac{k}{n})\frac{\lambda}{m}] \geqq 0$. Under the afore mentioned conditions, this inequality is evidently true. Requirement 2 reflects the monotony in the learning process. If the observed relative frequency

$\frac{i}{m}$ is equal to $\frac{k}{n}$ with $m > n$, then the value of $c(H, \frac{i}{m})$ should be closer to the observed value $\frac{i}{m}$ than $c(H, \frac{k}{n})$. If $\frac{i}{m} \leq \frac{k}{n} \leq \frac{1}{2}$ or $\frac{i}{m} \geq \frac{k}{n} \geq \frac{1}{2}$, this property should hold good a fortiori.

If $\frac{i}{m} \leq \frac{k}{n} \leq \frac{1}{2}$ and $n > m$, there is no conclusion possible about the ordering of the degrees of confirmation: The learning effect for n observations is stronger than the learning effect for m observations, so $c(H, \frac{i}{m})$ could have a value that is rather close to $\frac{1}{2}$ while $c(H, \frac{k}{n})$ could have a value rather close to $\frac{k}{n}$; this will happen for instance if n is large, m is relatively small and $\lambda \neq 0$. Some "natural" requirements can be defined only for special cases:

(3) $c(H, \frac{k+1}{n+1}) \geq c(H, \frac{k}{n})$.

For $\frac{k}{n} \geq \frac{1}{2}$, this inequality is a special case of requirement 2.

For $\frac{k}{n} < \frac{1}{2}$, it is natural to assume that one extra observation of H contributes positively to the confirmation that H will occur the next time. It can easily be verified that Carnap's confirmation function with (positive) constant λ, satisfies this requirement.

(4) $c(H, \frac{k+1}{n+2}) \geq c(h, \frac{k}{n})$ if $\frac{k}{n} \leq \frac{1}{2}$

$c(H, \frac{k+1}{n+2}) \leq c(H, \frac{k}{n})$ if $\frac{k}{n} \geq \frac{1}{2}$

This requirement describes the influence of two extra observations, once H and once T. Compare two situations, the first with the originally constructed confirmation function, the second in which a new confirmation function is constructed on the basis of a priori observation of H_2T_1. This new confirmation function can be defined as $c'(H, \frac{k}{n}) = c(H, \frac{k+1}{n+2})$. It has all the properties of a confirmation function, but clearly the a priori information will cause the weight of the observed relative frquency to be less than in the case $c(H, \frac{k}{n})$. Our open-mindedness to observations has decreased, because the insufficient reasoning is affirmed by H_2T_1. Again, this requirement is implicitly fulfilled by Carnap's confirmation function with constant (nonnegative) λ.

Let us now return to a learning process that is defined by some prior distribution $F(p)$, symmetric around $p = \frac{1}{2}$. It is not generally true that such a prior distribution generates a learning process that fulfills the afore mentioned requirements. Take, as an example, a prior distribution of which the density is given by

$$f(p) = 3 - 12p(1 - p)$$

Then we find

$$c(H, \tfrac{1}{3}) = \frac{\int_0^1 p^2(1 - p)^2(3 - 12p(1 - p))dp}{\int_0^1 p(1 - p)^2(3 - 12p(1 - p))dp} = \frac{2}{7}$$

and, apparently, this value is not in the interval $[\frac{1}{3}, \frac{1}{2}]$, as requirement 1 demands.

On the other hand, it can be shown that there are prior distributions that fulfill the above mentioned requirements, although they are not beta-distributed.

The following conclusions can now be drawn: (a) any prior distribution generates a learning process such that $c(H, \frac{k + 1}{n + 1}) \geq c(H, \frac{k}{n})$. This property is in agreement with the intuitively appalling idea that the extra observation of H leads to an increase of the degree of confirmation that H will occur at a new trial; (b) Carnap's confirmation function with constant value of λ is not the only possible confirmation function that fulfills the four requirements; (c) <u>if the four requirements are thought to be relevant, then any prior distribution that is not beta-distributed, should be checked on these four points</u>.

In Section 6 we shall meet examples where the requirements are not relevant, because the learning process is defined on the basis of other criteria.

5. THE NON-INFORMATIVE PRIOR

We now return to Carnap's confirmation function with constant λ. The value of λ reflects the relative weight given to the observed relative frequency $\frac{k}{n}$, if compared with the degree of confirmation $\frac{1}{2}$ given to the occurrence of the outcome H if no experience is available. For large λ, one is willing to change the value $\frac{1}{2}$ only if n is large and if $\frac{k}{n}$ differs considerably from $\frac{1}{2}$. For small λ, one is open-minded in regard to evidence even if this would imply large changes in the degree of confirmation that H occurs at the next trial after 1, 2, 3, ... observations respectively. According to Carnap, there is no reason to prefer one value of λ above another. It should be kept in mind, however, that in the confirmation function c(H, e) the experience e contains all relevant information available. This information is defined in observed relative frequencies. It is quite possible that other information is available in the form of knowledge about the structure of the experiment. If the experiment consists of coin tosses, then we shall adhere to a degree of confirmation of about $\frac{1}{2}$, and we shall deviate from it only after an exhaustive amount of observations. In other words, we shall choose a large value for λ. The choice of λ can be used for dealing with this kind of knowledge. At first sight, one is inclined to think that such a use of the λ parameter is inconsistent with the underlying assumptions of Carnap's learning model: all relevant experience should be contained in e, and the choice of λ should only reflect our attitude towards observed relative frequencies in general, independent of the specific experiment we are dealing with. It is possible, however, to define a learning process where λ has such a double interpretation, even within Carnap's set of assumptions. Consider for that purpose a pure Carnapian confirmation function

$$c(H, \frac{k}{n}) = \frac{k + \lambda . \frac{1}{2}}{n + \lambda}$$

Now suppose that at the start of the experiment, $2n_0$ trials have already been made, with n_0 times the observed outcome H. After n observations in the new sequence of trials, we have

$$c(H, \frac{k + n_0}{n + 2n_0}) = \frac{k + n_0 + \lambda.\frac{1}{2}}{n + 2n_0 + \lambda}$$

We may consider this as a new confirmation function, defined on the basis of a priori knowledge that consists of an observed relative frequency $n_0/2n_0$. The confirmation function on basis of this a priori knowledge can be written as

$$c'(H, \frac{k}{n}) = \frac{k + \lambda'.\frac{1}{2}}{n + \lambda'}$$

where $\lambda' = 2n_0 + \lambda$. Note that the value of λ increases, as a result of the a priori knowledge.

As we have already seen, there is a one-to-one correspondence between the confirmation function $c(H, \frac{k}{n})$ and the prior distribution $f(p) \propto p^{\frac{1}{2}\lambda-1}(1 - p)^{\frac{1}{2}\lambda-1}$. This prior distribution is dependent only on the learning parameter λ; it does not contain a priori knowledge. In the same way, the new learning function $c'(H, \frac{k}{n})$ generates a prior distribution $f'(p) \propto p^{\frac{1}{2}\lambda'-1}(1 - p)^{\frac{1}{2}\lambda'-1}$ in which the a priori information is contained, together with the originally defined learning process. The problem that rises is the following: If somebody constructs a (subjective) prior distribution, he defines implicitly both a priori knowledge and a learning process. The person should ascertain himself that his feelings are consistent as to both elements. This problem is touched upon by Raiffa and Schlaifer, when they state (9, page 61):

> "... it will usually be well to check subjective prior betting odds against hypothetical sample outcomes before beginning the actual analysis of the decision problem; and this in turn suggests that in some situations it may actually be better to reverse the procedure, making the intitial fit of the prior distribution agree with attitudes posterior to some hypothetical samples and then checking by looking at the implied betting odds."

In our terminology, subjective prior distributions should be checked for the learning process that is generated, and sometimes it is even better to define the prior distribution on the basis of the learning process. Raiffa and Schlaifer deal with learning effect in a slightly different way: they are mainly interested in the set of posterior distributions $f(p, \frac{k}{n})$ rather than the conditional probabilities $c(H, \frac{k}{n})$. This difference is not essential, however.

Up till now, we have dealt only with prior information of the form of relative frequencies $n_0/2n_0$, symmetric as to the outcomes H and T. One could easily generalize this to frequencies k_0/n_0. For that purpose, we introduce the notation m(H) for the degree of confirmation that H occurs at a new trial if no experience is available. Symmetry arguments lead Carnap to the assumption $m(H) = m(T) = \frac{1}{2}$. Now suppose that a confirmation function is constructed on the basis of a priori knowledge of an observed relative frequency k_0/n_0. The new confirmation function is related to the pure Carnapian learning function by

$$c'(H, \frac{k}{n}) = c(H, \frac{k + k_0}{n + n_0}) = \frac{k + k_0 + \lambda.\frac{1}{2}}{n + n_0 + \lambda}$$

Writing m'(H) for $c(H, \frac{k_0}{n_0})$, we can easily see that $c'(H, \frac{k}{n})$ can be written as

$$c'(H, \frac{k}{n}) = \frac{k + \lambda' m'(H)}{n + \lambda'}$$

with $\lambda' = n_0 + \lambda$. The prior distribution corresponding to the new confirmation function is

$$f'(p) \propto p^{\lambda' m'(H)-1}(1 - p)^{\lambda'(1-m'(H))-1}$$

Again, the prior distribution is beta distributed. For $k_0/n_0 \neq \frac{1}{2}$, it is not symmetric.

The formal definition of the confirmation function $c'(H, \frac{k}{n})$ does not violate the underlying assumptions of Carnap's theory. It is derived directly from the "pure" confirmation function $c(H, \frac{k}{n})$. It is important in that it paves the way to a generalization of Carnap's theory. One of the serious drawbacks of that theory is the restriction that any statement should be in terms of frequencies. There is, however, a considerable amount of knowledge about phenomena that is not in the form of frequencies. If one is able to transform that kind of knowledge into validations of m'(H) and the relative weight given to this value in comparison to observed relative frequencies (the validation of λ'), then a confirmation function $c'(H, \frac{k}{n})$ can be constructed in an analogous way. The validation may be subjective, and in that case the confirmation function has subjective elements. The only difference with the subjectivistic model is the learning parameter λ that is implicitly

present in both λ' and $m'(H)$. If someone constructs a prior distribution $f'(p)$, is he then evaluating his learning process or his uncertainty about some "true" value of p? Or are these two things essentially the same? These questions have yet to be answered.

The four requirements for a learning process, as given in Section 4, were related to symmetric situations as to the outcomes H and T. This symmetry is disturbed in the new confirmation function $c'(H, \frac{k}{n})$. It can easily be verified that analogous requirements are fulfilled for constant λ':

(1a) The value of $c'(H, \frac{k}{n})$ is always between $m'(H)$ and $\frac{k}{n}$;

(2a) If $\frac{i}{m} \leq \frac{k}{n} \leq m'(H)$ with $m \geq n$, then $c'(H, \frac{i}{m}) \leq c'(H, \frac{k}{n})$

(3a) $c'(H, \frac{k+1}{n+1}) \geq c'(H, \frac{k}{n})$

(4a) $c'(H, \frac{k+rt}{n+t}) \geq c'(H, \frac{k}{n})$ if $\frac{k}{n} \leq m'(H)$ and $m'(H) \leq r \leq 1$

where $k + rt$ is an integer.

The arguments for these restrictions are the same as have been mentioned in Section 4.

Now that we have seen that the parameter λ' contains both the original learning parameter λ and the relative weight n_0 that is given to the a priori information $m'(H)$, we might ask ourselves whether it is possible to limit the range of acceptable λ values. Such λ values can be used for the construction of "pure" confirmation functions $c(H, \frac{k}{n})$ where no prior information is available. The corresponding prior distributions $f(p)$ could be called non-informative priors.[1]

---

[1] Carnap favours the confirmation function c^* that gives a priori weight to all so-called "structure-descriptions" of samples with a fixed number of trials. A structure-description is defined by the relative frequency with which H occurs independent of the individual outcomes. The function c^* is characterized by the corresponding function m^* with

$$m^*(H_2H_1) = m^*(H_2T_1 \text{ or } T_2H_1) = m^*(T_2T_1) = \frac{1}{3}$$

$$m^*(H_3H_2H_1) = m^*(H_3H_2T_1 \text{ or } H_3T_2H_1 \text{ or } T_3H_2H_1) =$$

$$= m^*(H_3T_2T_1 \text{ or } T_3H_2T_1 \text{ or } T_3T_2H_1) =$$

$$= m^*(T_3T_2T_1) = \tfrac{1}{4} \text{ etc.}$$

If r_n denotes the relative frequency in a sample of n trials, we have

$$m^*(r_n = 0) = m^*(r_n = \frac{1}{n}) = \ldots = m^*(r_n = \frac{n}{n}) = \frac{1}{n}$$

and for n tending to infinity, we obtain the rectangular prior. If there are only two possible outcomes for each trial, the function c^* is characterized by $\lambda^* = 2$.

Carnap believes that all positive values of λ are acceptable. Subjectivists often use a rectangular distribution for the non-informative prior, and this corresponds to $\lambda = 2$.[1] We believe that $\lambda = 0$ is the only acceptable choice.

Let us first consider the case that $\lambda = 2$ is chosen for the non-informative prior. The reason for this choice is the principle of insufficient reasoning, applied to the set of possible p-values. It is oftern argued that the principle cannot be used for a case like this: instead of p, we could use some non-linear transformation of p as parameter in the binomial distribution, and if the principle of insufficient reasoning is applied to this new parameter, a different result will be obtained. On the other hand, one might argue that p is not just a parameter, but a "natural" quantity that can be interpreted as a relative frequency in a very long (infinite) sequence of observations.

One might well apply the principle of insufficient reasoning to some interpretable quantity, where uninterpretable non-linear transformations are thought to be irrelevant. We consider this defense to be rather doubtful: what is "interpretable"? Taking the learning process into consideration, it is hardly necessary to discuss this point. There is no apparent reason for preferring $\lambda = 2$ above any other value of λ close to it. This leaves us with only two possibilities: Either all values of λ are acceptable, or one of the extreme values of λ should be chosen.

The "pure" confirmation function can be written as $c(H, \frac{k}{n}) = \frac{k + \lambda.\frac{1}{2}}{n + \lambda}$. For $\lambda = \infty$, the degree of confirmation is $\frac{1}{2}$, independent of any observations. This indeed would be a very unfortunate learning function. For $\lambda = 0$, the degree of confirmation is completely determined by the observed relative frequency. In the latter case, we are completely open-minded as to new data. For any $0 < \lambda < \infty$, we take into consideration the observed relative frequency, but at the same time we adhere to the value $\frac{1}{2}$ to a certain extent. In a non-informative situation, the value $\frac{1}{2}$ is derived from the <u>logical</u> symmetry between the two possible outcomes. It reflects our lack of knowledge and it gives no factual information. But then, it does not make sense to define the degree of confirmation after the observation of $\frac{k}{n}$ as a convex combination of the factual information $\frac{k}{n}$ and the reflection of ignorance $\frac{1}{2}$. Ignorance should have no weight in comparison with some well-defined amount of knowledge. This line of reasoning suggests that $\lambda = 0$ should be chosen.

The same kind of argument can also be expressed as follows: in a non-informative situation there is no knowledge available. But then, evidently, we should be as maximum open-minded as possible as to data, this being the only kind of information that is available. The choice $\lambda > 0$ is feasible, but not plausible in a context of a non-informative situation.

The choice $\lambda = 0$ corresponds to a prior distribution with half of its mass on the point $p = 0$ and half of its mass on $p = 1$. This is why Raiffa and Schlaifer reject the choice $\lambda = 0$; they argue as follows (9, page 65):

> " (...) we find that the beta distribution does not approach a proper limiting distribution: namely a two-point distribution with a mass of m' on $p = 1$ and a mass $(1 - m')$ on $p = 0$.[1] Now this limiting distribution cannot in any sense be considered "vague". On the contrary, it is completely prejudicial in the sense that no amount of sample information can alter it to place any probability whatever on the entire open interval $[0, 1]$. A single sample success will annhilate the mass at $p = 0$, and a single failure will annihiliate the mass at $p = 1$; but a sample containing both successes and failures will give the meaningless result 0/0 as the posterior density at all p in $[0, 1]$ and also at the extreme values 0 and 1 themselves."

The argument is proceeded by the consideration of a beta distribution with λ close to 0. Such a distribution concentrates nearly all the probability mass close to the points $p = 0$ and $p = 1$, and

> "it requires a very great deal of information in the ordinary sense of the word to persuade a reasonable man to act in accordance with such a distribution even if the probability assigned to the interval is not strictly 0. Long experience with a particular production process or with very similar processes may persuade such a man to bet at long odds that the fraction defective on the next run will be very close to 0 or very close to 1, but he is not likely to be willing to place such bets if he is completely unfamiliar with the process."

---

[1] In our case, $m' = 1 - m' = \frac{1}{2}$. The underlinings are part of the quotations.

The first argument is inappropriate: if we define the posterior distribution with the aid of a prior distribution then the undefined expression 0/0 could be given a meaning by taking the limit $\lambda \to 0$ in the prior distribution as well as in the posterior distribution. If the learning process is defined explicitly as in the Carnap situation there will be no problems at all!

The second remark deserves more attention. The example chosen by Raiffa and Schlaifer is not a very fortunate one. Nearly all production processes will generate good products and defective ones, so it is known beforehand that one is dealing with some stochastic phenomenon. The situation cannot be called non-informative. A real non-informative situation is the one for which no analogous situations exist. This implies that we do not know whether a deterministic model or a stochastic model is appropriate. It is not at all unrealistic to start with the idea that the observed phenomenon is deterministic. Such an attitude implies a prior distribution with its probability mass on its point $p = 0$ and $p = 1$. The stochastic model seems appropriate only after the observations H and T; the probability mass of the posterior distribution will be spread over the complete interval $0 < p < 1$. The argument made by Raiffa and Schlaifer, that one should be very convinced, before one puts all probability mass on the points $p = 0$ and $p = 1$, is misleading.

It is not <u>necessary</u> to have much information pointing to a value of p close to either 0 or 1, in order to use a prior distribution with λ close to 0. On the other hand, if such information is available, it is a <u>sufficient</u> reason to use this prior distribution. If this prior distribution is objectively determined, the pair of observations (H, T) will be (almost) impossible to obtain. The posterior distribution after this pair of observations is then irrelevant.

Raiffa and Schlaifer are right, when they state that long experience with a particular production process may persuade a man to bet on "p is either 0 or 1"; it is not true, however, that the man is not likely to bet in this same way if he is completely unfamiliar with the phenomenon (better than: production process) in question.

It appears that there is an essential difference between the situation where the prior distribution is objectively determined, and the situation where the prior distribution is derived from the learning process. In the next section we shall discuss this matter in more detail.

6. THE INTERPRETATION OF PRIOR DISTRIBUTIONS

Up till now, we have carefully avoided drawing conclusions as to the interpretation of prior distributions. On the one hand, if the learning process is defined by means of a confirmation function, the prior distribution can be considered as a direct consequence of this learning process; f(p) describes betting odds for compounded bets on relative frequencies for large numbers of (future) observations; the bets are called compounded, because the hypothetical outcomes of the first trials influence the confirmation about later outcomes in a way that is described by the learning process. On the other hand, there could be some collective or random devices analogous with the particular device that is used for the (future) sequence of trials; f(p) can then be interpreted as a distribution on "true" p-values; it describes betting odds on relative frequencies for large numbers of future observations; such an objective prior generates a learning process. If no a priori information is available, then we know nothing of collectives and we are forced to define a learning process. If some a priori knowledge is available, it could influence the learning process directly, but it could also be related to some collective and we might then choose the determination of a prior distribution as the best way of incorporating this kind of information in the learning model.

In the following examples, we describe the way in which we personally would like to express the a priori knowledge so as to let it fit into the model.

(1) Assume that some newly developed chemical liquid is put into a goldfish bowl. The question is whether or not the goldfish will survive. In this example it is assumed that no experience is available with analogous chemical liquids, and that the reactions of fish to this liquid cannot be derived from other properties of the liquid that have been studied before. Under such circumstances, it is impossible to define a relevant collective to which this liquid belongs. We are in a typical non-informative learning situation. Take $\lambda = 0$, and the corresponding prior distribution is the one with half its probability on $p = 0$ and half of its mass on $p = 1$, in other words, it is assumed that all fish will either survive or die.

(2) In a psychological institute, a person is asked to determine degrees of belief on outcomes H, after observation of a sequence of outcomes. He is told that the outcomes are generated by a binomial random process. The "true" value of p is not known to him. Evidently, this person has no idea about a collective of p-values, part of which is the specific "true" p-value. He therefore has to construct a learning function. He will not take $\lambda = 0$, because he has the information that the real process is a random one. To be open-minded in regard to data (the only kind of knowledge that he will get), he will choose a low value of λ, for instance $\lambda = 1$. The choice $\lambda = 2$ is equally possible but by no means necessary. As a matter of fact, people seem to be inclined to choose very low values of λ, even negative ones. As soon as $\lambda < 2$, is chosen, one calls the person "conservative". The behaviour of people is compared with the behaviour that would correspond to a rectangular prior distribution. If people are more open-minded to data than that, it is thought that they are conservative as to the information they have. This phenomenon of conservatism is often observed, even in situations where objective prior distributions are given to the subjects. In the example given here, there is no objective prior, and people are fully justified in being more open-minded to data than the rectangular prior would admit. The phenomenon of conservatism should be measured against the non-informative prior corresponding to $\lambda = 0$, instead of the wrongly defined so-called "non-informative prior", corresponding to $\lambda = 2$. The results of Edwards [3] seem to point at conservatism even when compared to $\lambda = 0$, although less dramatically than described by Edwards, who has teken the rectangular prior as the non-informative one.

(3) A symmetric disk, the surface of which is 2/3 red and 1/3 green, is spun around. It will come to rest on one point of its edge. If the outcomes red and green are notated by R and G respectively, it seems to be sensible to assert $m(R) = 2/3$ and $m(G) = 1/3$. It seems to be very artificial to assume a collective of disks with different "true" p-values, to which the specific disk belongs. Instead, we shall

the learning process. The rather strong evidence in the mechanical sense will lead us to a rather high value of λ. The confirmation function has to be tested for the influence of several hypothetically observed relative frequencies, for instance $c(C, G)$; $c(G, \frac{3}{3})$; $c(G, \frac{5}{5})$; ... λ should be chosen such that these values are acceptable to us.
If the disk is not symmetric, we could still believe

that m(G) = 1/3. The mechanics of the experiment are less clear, and a lower value of λ will be chosen.

(4) The experiment consists of coin tosses, without previous experience with the specific coin involved. Evidently, one may choose m(H) = $\frac{1}{2}$. Now it is possible to continue in the same way as described under (3). Symmetry reasons could lead to the choice of a (high) value of λ. In this case, however, it also seems possible to represent our a priori knowledge about coins in general in the form of a prior distribution on the collective of coins with different "true" p-values. A man who has much experience with coin-tossing experiments, with a great number of different coins, will know the proportion of coins of which the relative frequency has stabilized on p = r for different values of r. In other words, he knows the distribution f(p). When he uses this distribution as the prior distribution he is forced into a learning process that is not necessarily the same learning process as the one that he has derived from the mechanical properties of the specific coin in question.

Which one of the two learning processes should be preferred? First, the person should check as to whether in his opinion the coins in the total set are comparable with the specific coin that is used for the coming sequence of trials. If L_s, $f_s(p)$, L_c and $f_c(p)$ denote the learning process and prior distribution based upon the single coin and based upon the collective of coins respectively, then he might check as to whether he would choose L_s for all coins in the collective. If the answer is in the negative due to the fact that he feels that there is a difference between a quarter and a dime for instance, then the collective should be restricted to comparable coins. Under the assumption that the collective consists of comparable coins, it remains possible that L_s and L_c are different. Now if L_s is used to determine betting quotients on relative frequencies in very long sequences of observations for all coins in the collective, then these betting quotients would be obtained from $f_s(p)$, and experience has taught that $f_c(p)$ gives better results. Now that we know $f_c(p)$, it is better to adapt the learning process to this kind of knowledge. Therefore, L_c seems to be a better choice than L_s. The objectively determined prior is preferred to the subjective evaluation of the mechanics of the experiment with a single coin.

It is quite possible that the prior distribution $f_c(p)$ violates some of the requirements for learning processes, as developed in Section 5. In the case that $f_c(p)$ is objectively determined, this appears to be quite acceptable. Let us compare the two learning processes L_s and L_c. For L_s, we had a priori knowledge m(H), where $m(H) = \frac{1}{2}$ if H and T are symmetric. After the observation of a relative frequency $\frac{k}{n}$, there is reason to deviate from m(H); there is, however, no reason to deviate further than this value $\frac{k}{n}$. For L_c, we have a priori knowledge of $f_c(p)$; the specific coin is considered as one of the many coins with different "true" p-values. The observation of $\frac{k}{n}$ successes makes us reconsider the probability that the specific coin has a "true" probability p, i.e., we may calculate $f_c(p/\frac{k}{n})$. This might result in a probability (rather than a degree of confirmation) $c(H, \frac{k}{n})$ that is not in the interval $[\frac{k}{n}, \frac{1}{2}]$, or $[\frac{1}{2}, \frac{k}{n}]$.The first requirement can be violated, as has been shown in Section 4, and, in the same way, other requirements can also be violated.

(5) Now take the same example as under (4). In this case the person has no objective knowledge about the collective of coins. He might still construct the collective as a hypothetical device. Very vague knowledge about other coins might lead him to the construction of a prior distribution $f_c(p)$. On the other hand, the learning process L_s generates a prior distribution $f_s(p)$. Formally, both prior distributions are treated in the same way. The conflict between those two prior distributions can be solved only by trying to find a prior distribution that fits both purposes. If, for instance, one prefers a learning process L_s where one is completely open-minded as to new data, so $\lambda = 0$, then one is forced to bet upon relative frequencies in large sequences of observations by giving equal chances to the outcomes "always H" and "always T". Other possibilities are excluded. Now this is a kind of a bet that is unacceptable in view of our experience with coins in general. On the other hand, the knowledge about the collective is so vague, that we shall not use a prior distribution $f_c(p)$ that violates the requirements of a learning process. Under these vague circumstances both L_s and $f_c(p)$ can easily be changed: our intuition lacks precise to such a degree that changes to a rather wide extent are acceptable.

We now summarize our findings.

- Both Carnap's confirmation function and the subjective approach based upon some prior distribution $f_c(p)$ describe the way in which we learn from experience. They must not be compared with the frequentistic probability theory, but with the inference rules used in the application of the frequentistic theory to finite problems.
- Formally, Carnap's theory can easily be extended such that subjective a priori knowledge can be incorporated in the model. As a mathematical extension, a prior distribution $f_s(p)$ can then be introduced, where p denotes the relative frequency of successes in large (future) sequences of observations, and where $\int_\Delta f_s(p)$ denotes degrees of confirmation for intervals Δ of possible p-values. These degrees of confirmation can be considered as compounded betting stakes, where the (future) first n outcomes will give a learning effect for later outcomes. This makes sense even in the situation where no collective of analogous experiments with different "true" p-values can be imagined.
- If a collective with different "true" p-values exists, one may construct a prior distribution $f_c(p)$, and use this distribution for bettings on relative frequencies in long runs of trials with a device with some specific p-value. In that case, a learning process is forced upon us, and the requirements for a learning process as derived for the extended Carnap model can be violated.
- The non-informative prior cannot be derived from such a collective, because, by definition, we have no information about the collective. Therefore, the non-informative prior is defined on the basis of the learning process, and we argue that maximum open-mindedness as to data is the best choice to be made. The rectangular prior distribution is not non-informative, and the introduction of it is due to the (wrong) interpretation of $f_c(p)$, instead of the correct interpretation of $f_s(p)$.

- The assumption that we can learn from observed relative frequencies, independent of the order of the observations, implies that we believe in something like stabilizing relative frequencies. It would not make sense to use the observed relative frequency in the first 100 observations for a degree of belief for the 10,000st outcome if we did not think that the relative frequencies would stabilize. It is not relevant, however, whether or not the relative frequencies stabilize in reality. We act in accordance with our best judgment, but nevertheless, in reality, our judgment may be wrong. It is impossible to find out what happens in infinite long sequences of trials, so we can never prove that our judgment is wrong. We could, however, find some indications for it in a long though finite sequence. Such indications, however, would involve ordering numbers of the trials. Our a priori assumption that the sequence of events is symmetric excludes the possibility of reacting upon indications of this kind. In some way or other, the subjective probability model should be extended with testing procedures that permit us to discard the model if the underlying assumptions appear to be unrealistic.
- The pure logical approach of Carnap gives rise to some problems. First, we must define our experiment carefully. If H denotes "having a tail", the relative frequency will decrease according as the objects being tested move from animals to animals including human beings, or to any object whatever. Furthermore, we must define the attribute that is tested. These two definitions require at least some a priori knowledge. But then, it is open to question as to whether the concept of a non-informative situation makes sense. We have already seen that the example of a non-informative situation as given in Section 6, was very artificial. This objection favours the extension of Carnap's model to a similar learning model into which subjective ideas can be fitted. The non-informative situation should be considered as a mathematical idealization, that does not exist in reality, but that is approximated quite closely for some experiments. In the same way, one could consider stabilizing relative frequencies in infinitely long sequences of observations as a mathematical idealization that does not exist in reality, but that can be approximated. The non-informative situation and the "true" value of p are limiting points on a scale of observed relative frequencies, where only the intermediate points are relevant for realistic problems.

REFERENCES

[1a] Carnap, R. Logical Foundations of Probability, The University of Chicago Press (1950).

[1b] Carnap, R. Induktive Logik und Wahrscheinlichkeit, (Bearbeitet von W. Stegmüller), Springer, Wien (1959).

[2a] Dantzig, D. van, "Statistical Priesthood (Savage on Personal Probabilities), Statistica Neerlandica 11, pp. 1-16 (1957).

[2b] Dantzig, D. van, "Statistical Priesthood II (Sir Ronald on Scientific Inference), Statistica Neerlandica 11, pp. 185-200 (1957).

[3] Edwards, W., Conservatism in Human Information Processing in B. Kleinmuntz (ed.), Formal Representation of Human Judgment, Wiley, New York (1960).

[4] Feller, W. An Introduction to Probability Theory and its Applications, Wiley, New York (1957-1966).

[5] Finetti, B. de, "La Prévision: Ses lois logiques, ses sources subjective, Ann. de l'Inst. Henri Poincaré 7, pp. 1-68 (1937).

[6] Fisher, R., Statistical Methods and Scientific Inference, Oliver and Boyd, Edinburgh/London (1956).

[7] Leede, E. de and J. Koerts, "On the Notion of Probability, A Survey", Report 7007, Econometric Institute, Rotterdam (in print in Methodology and Science).

[8] Mises, R. von, Mathematical Theory of Probability and Statistics, Academic Press, New York/London (1964).

[9] Raiffa, H. and R. Schlaifer, Applied Statistical Decision Theory, Harvard University, Boston (1961).

[10] Savage, L.J., The Foundations of Statistics, Wiley, New York (1954).