Stat.

(Netherlands school of economics)

ECONOMETRIC INSTITUTE

ON THE NOTION OF PROBABILITY: A SURVEY

E. de LEEDE and J. KOERTS

Erasmus

REPORT 7704/ES

ERASMUS UNIVERSITY ROTTERDAM, P.O. BOX 1738, ROTTERDAM, THE NETHERLANDS

ON THE NOTION OF PROBABILITY

A SURVEY

By E. DE LEEDE AND J. KOERTS

Abstract

Different interpretations of the probability concept are considered. In a situation of repeated trials, indepedent in the frequentistic sense, the frequentistic theory appears to be adequate only if a sufficiently large number of observations is available. In those cases where only a limited number of observations is available, the subjectivistic approach towards probability might be of help. The underlying assumptions of this notion of probability are considered. From the fundamentals of normative probabilities, the ambiguity of a prior distribution for an "unknown" parameter is shown: Sometimes such a prior distribution can be considered as a reflection of (personal) beliefs, sometimes it must be considered as a derivative of the learning process that is chosen. When these two interpretations overlap, the meaning of a prior distribution becomes rather vague.

Contents

This report is based on report 7007 of the Econometric Institute; it is completely rewritten.

We added a section "On fair bets", according to the ideas of Professor de Finetti and rewrote Carnap, according to his last ideas which became known after his death.

Moreover we introduced a section Bayesianism; a mixed approach and striked the section of B.O. Koopman.

Finally, we added the concluding remarks.

1. Introduction

Mathematical statistics has shown a strong development over the past two centuries. That is, the number of techniques for analysing data has increased enormously. These techniques are all based on probability theory and they have been developed so as to end up with some kind of inference, some statement about the real world. The concept of probability, however, is rather abstract and hence the question arises: what is the interpretation of the notion of probability in real world situations?

In answering this question we reach the foundations of statistics and, as is the case in many sciences, it is precisely here that differences in opinion appear. The foundations of this concept (and hence the foundations upon which all techniques are based) have been the object of much discussion . Different interpretations have resulted in different axiomatic systems. Unfortunately, however, not all systems justify all the techniques already in use. Thus research workers using statistical methods should pay attention to the foundations of the notion of probability. Economists in particular have reason to be interested in views about these because their statistical problems differ quite often from the assumptions made for the several theoretical concepts of probability: The frequentistic theory may be well equipped for applications in experimental sciences, in situations where the number of observations can be chosen arbitrarily large, but the economist is most often faced with problems where no experimenting is possible. The subjectivistic approach, on the other hand, may be useful for dealing with decision problems with a limited amount of information, but the science of economics is traditionally of a normative nature and subjective elements do not fit well into this picture.

About 1930, the problem of interpretation of the probability concept seemed to be solved by the introduction of an abstract axiomatic system as the basis of probability theory; see for instance Kolmogoroff [10]. Probability theory now became a part of mathematical measure theory. This approach was important because the <u>probability calculus</u> could be developed without being hampered by problems of interpretation. But, as soon as one tries to apply statistical methods, one must confine oneself to a particular interpretation of the notion of probability! This interpretation then determines the techniques available to the research worker as well as the area in which statistics can be used.

The economist now runs into a problem because the literature is extensive, very technical, and therefore difficult to grasp.

The purpose of this paper is to present a brief survey of several axiomatic systems, which correspond to different interpretations of the concept of probability. We shall discuss their strengths and weaknesses and try to answer the question: "What should the attitude of an econometrician towards the notion of probability be?"

It is not our primary purpose to join the discussion about rather philosophical issues, like "could a claim to so-called objectivity be justified in any sense", or "could subjective concepts be incorporated in a scientific theory, without destroying the scientific character of the theory". This kind of questions are of minor importance for the practical research worker if compared with questions like "what are the practical limitations of either theory, and under what kind of circumstances is the use of one system or the other permitted". In summarizing the several axiomatic systems, we shall mainly concentrate on those issues that are relevant from this rather practical point of view.

2. The Classical Approach

The theory of probability originates from gambling problems. In such situations the following definition was used implicitly for a long time and later defined explicitly by Laplace:

D.C.1. The probability of an event occurring is equal to the ratio between the number of cases favourable to this event and the total number of possible cases, provided that all these cases are mutually symmetric.

This definition depends on the notion of symmetry and so raises a logical problem: if probability is defined on the basis of symmetry, how should we define this symmetry? The problem can be illustrated with an example. Consider the throw of a die. In this case we all believe that the six faces are symmetric. But what does that mean? If we consider the faces in more detail we realize that each is different, if only in its identifying marks. In this sense there is no symmetry. The die could, however, be called symmetric in its relevant (mechanical) aspects. But if our experiment were to consist of taking the die from the table and returning it with the same face up, then the six events would apparently not be symmetric. Therefore, the six events should be called symmetric only in those aspects relevant to some specified experiment. But this means that the six events should be equally probable, because equal probabilities is exactly the aspect that is thought to be relevant. So, mutually symmetric cases imply equal probabilities and this makes the definition circular. This fact is well-known, but it is perhaps less well-known that Jacob Bernouilli [1] had evaded this problem by a different definition.

D.C.2. Probability is a degree of belief (gradus certitudinis). A degree of belief is measured by the ratio between the number of favourable cases and the number of possible cases if all cases are thought to be equally likely.

In this definition, a distinction is made between the contents of the probability concept and the way in which such probabilities should be measured. This distinction is of vital importance and will be met again later.

Are all problems solved if we restrict ourselves to gambling cases only, and if we accept Bernouilli's interpretation? The answer is negative for, even in this case, the concept of equally likely cases raises difficulties. Consider two coin tosses. Let H represent heads and T tails. Traditionally the events (HH), (HT), (TH), (TT) are thought to be equally likely. However, another description of the possible outcomes suggests another kind of symmetry. We could just as well assume that the outcomes no heads, heads once and heads twice are equally likely. This view point has been advocated by no less a person than d'Alembert. The difference in opinion leads, of course, to different numerical probabilities for the same event; and in gambling this is unsatisfactory.

It has been shown that, even for one person, symmetry arguments can lead to sets of equally likely cases that are mutually contradictory. This is more serious for it means that, even in the opinion of a single person, equally likely cases are not uniquely defined and thus it makes no sense to define the measurement of probabilities in terms of the many sided concept of equally likely cases. The following example illustrates the problem.

Consider a car that needs 1 liter of fuel for some distance between 8 and 12 kilometers. Symmetry arguments would lead us to a 50% chance that the actual efficiency is between 1 liter for 8 and 1 liter for 10 kilometers. Often the efficiency is measured by the number of liters needed for a distance of 1 kilometer. In our case, this value should be between 0.125 and 0.083 liters. The same kind of symmetry arguments would lead us to a 50% chance that the actual efficiency is between 0.125 and 0.104. These values correspond to 1 liter for 8 kilometers and 1 liter for 9.6 kilometers and contradicts the values found earlier.

Thus we can say:

(1) Laplace's definition is circular. Jacob Bernouilli solved this problem by distinguishing between the concept (degree of belief) of probability and the way in which it is measured.

(2) The measurement of probability is only defined if equally likely cases can be determined. This implies that many empirical phenomena which show statistical regularity are excluded. Such a definition is too restrictive and therefore can not be accepted.

(3) Even if equally likely cases can be determined, the measurement of our degree of belief is not necessarily unique.

3. The Frequency Approach

As has already been said, we cannot accept the classical definition of probability because its scope is far too limited.

Since 1850 there have been repeated attemps to introduce a suitable definition of the notion of probability. Such a definition should serve two purposes: (1) it should extend the field of the applicability of probability theory and (2) it should solve the problems met in the previous section.

Von Mises [15,16] made an important contribution towards a solution. He can be regarded as one of the founders of the frequency approach in probability theory in which we can say that the object of the theory is to provide a mathematical model suitable for the description of a certain class of observed phenomena known as random experiments. Such experiments are characterized by the following properties:

(a) The outcomes differ from trial to trial.

(b) The relative frequency of observed events stabilizes for increasing numbers of observations.

(c) There is no such thing as a successful gambling system. In this, a gambling system is to be understood as a system that prescribes bets on basis of, for instance, the ranking number of the trial at hand.

Successful betting schemes are possible if the scheme would adapt itself to a bias of the roulette wheel, but this is not the kind of gambling systems that Von Mises had in mind. In the latter case, the gambler "learns" the probability value during the game, whereas (c) claims the absence of gambling systems in situations where the probability values (here: stabilized relative frequencies) are fully known.

The properties (b) and (c) give rise to substantial logical questions. How can we know that relative frequencies stabilize in the long run? Does it make sense to speak of the absence of gambling systems in situations where the probability values are fully known, if it would take an infinite number of observations to know these values? Even if the stabilized relative frequencies would be known, how can we prove the non-existence of successful gambling systems? The frequentist's answer to these questions will be that, first, there can be no objection against the building of a mathematical model that fulfils (a) through (c); second, there are many situations where relative frequencies tend to stabilize in long though finite sequences of observations and where gambling systems have broken down; third, the application of the mathematical model has proved to be successful in many cases; and fouth, there is not yet another model that describes such phenomena better. To understand the frequentist's attitude, we should distinguish carefully between reality (our experience of stabilizing frequencies and failing gambling systems), the mathematical model that is constructed, and the problems that arise when the model is applied to real situations. The mathematical model should reflect the properties that we meet in reality, and a terminology has to be introduced that makes the requirements (a) through (c) more explicit.

A first basic concept of the model is the notion of a label space.

D.M.1. The label space S is the set of all conceivable outcomes of an experiment.

Many authors use the term sample space for this concept. As has already been said, in reality we deal with long (though finite) sequences of observations. The theoretical counterpart in the model of such a sequence of observations is an infinite sequence of labels $\{x_j | j = 1, 2, \dots\}$ where each x_j is a point of the label space S. Again in reality we experience, in a long sequence of trials, that the relative frequency of occurrence of a given outcome becomes more or less constant. The counterpart of this fact is the first basic assumption that, in an infinite sequence $\{x_j\}$, the relative frequency of each label a_i tends to a fixed limit p_i:

D.M.2. If in an infinite sequence of labels $\{x_j | j = 1, 2, \dots\}$ the relative frequency of each label a_i tends to a fixed limit p_i, then this limit is the <u>chance</u> of a_i within the given sequence.

It may be noted that an infinite sequence, consisting of alternating labels 0 and 1, fulfils the requirements for definition 2. Such a sequence is completely structured, and a successful gambling system is easily derived. For an explication of randomness, Von Mises introduces the concept of place selection.

D.M.3.: A place selection is a method of selecting an infinite subsequence out of an original sequence which satisfies the following requirements: The decision to retain or reject the n^{th} element of the original sequence may depend on the number n and on the label values $x_1, \ldots, x_{n-1}$, of the n-1 preceding elements but never on the label value of the n^{th} element or any following element.

One would be inclined to reserve the word "probability" for those cases where chances are invariant for <u>any</u> place selection. A simple example can make it clear that this is not possible: Think of an infinite sequence of the labels 0 and 1 that is generated by flipping a fair coin. The chance of label 0 is $\frac{1}{2}$. There is in the set of all place selections a place selection that happens to consist of only labels 0. Although the original sequence was meant to be a random sequence, the chance of label 0 changes from $\frac{1}{2}$ to 1 for this specific place selection. Apparently, we have to be more modest in our requirement of invariance. The word "probability" will be used for chances that are invariant for a system of infinite denumerable place selections. Formally, we define:

D.M.4. Let S be a discrete label space and $K = \{x_j\}$ an infinite sequence of elements of S. Let G be a system of denumerably many place selections. We assume that

(1) for every label a_i in S the limiting frequency p_i exists in K;

(2) $\sum_{j=1}^{\infty} p_i = 1$, the sum extended over all elements of S;

(3) any place selection Γ belonging to G applied to K produces an infinite subsequence of K, in which again, for every a_i, the limiting frequency exists and is equal to p_i. Then, K is called a collective with respect to G and S, and p_i is the probability of a_i.

Within the collective, several operations can be considered, alone or in combination. These operations lead to new collectives and new probabilities. Four elementary operations can be distinguished:

(1) Selection: a new sub-sequence is derived from an original sequence by means of a place selection. This operation leads to a new collective. Probabilities are not changed by it.

(2) Mixing: this is the reconstruction of the collective by taking several labels together. For instance, a sequence consisting of the numbers

1 through 6 is changed by only discribing the original elements as even or odd. Probabilities change in the sense of the usual addition rules.

(3) Partition: from the original collective, a new one is formed by selecting only those elements that satisfy some specified condition. Probabilities change in the sense of conditional probabilities.

(4) Combination: from two collectives with labels a_i and b_j, we can define a new collective with the pairs (a_i, b_j) as its labels. Probabilities change corresponding to the product rule.

In this theory, probability calculus is nothing more or less than the art of going from one collective to another with the aid of the four elementary operations. A probability problem usually has the following structure: given a collective and probabilities in it, find the probabilities in another well-defined collective.

It is perhaps good to state here explicitly that Von Mises succeeds in building up a probability calculus for both discrete and continuous label space. His results for the continuous case do not, however, coincide with the usual axiomatic approach. To be more precise: in the case of a continuous label space, the class of events to which probabilities are assigned is not a sigma algebra. If for instance we take the closed interval [0, 1] for our label space, probabilities are only assigned to those subsets of [0, 1] which are Jordan sets instead of all Borel sets as is usually done. This difference originates from the fact that Von Mises introduces an extra requirement namely: _a probability statement should be verifiable, at least conceptually_. Given this requirement, the class of Borel sets obviously cannot be used. We shall now discuss some of the difficulties associated with the frequency approach.

It is good to quote Jeffreys [9] at this point. In his opinion, "a definition is useless unless the thing defined can be recognized in terms of the definition when it occurs. _The existence of a thing or the estimate of a quantity must not involve an impossible experiment_."

If we accept this principle, Jeffreys' objection is very serious because the notion of probability as defined by Von Mises is completely hypothetical, being a limit in an infinite sequence. It is an abstraction and can never be verified!

We must realize, however, that essentially hypothetical models are used in all branches of science. Models are always abstrations of reality. It is therefore not necessary for their variables to exist in reality, at least in our opinion. Euclidean geometry, for instance is such a model; its notions such as "points" and "straight lines" and "triangles" do not exist in reality and can never be verified. Yet the model has proved to be very useful. Thus we do not accept Jeffreys' principle. As to the question of verifiability, we should distinguish between verifiability of concepts introduced within the model (such as points, lines, and triangles) and verification of conclusions from the model, that is, the applicability of the model. Concepts like probability introduced within the model need not and usually cannot be verified because they are abstractions. In our opinion the only requirement for a model is that the axioms on which it is based be consistent. In this respect, Von Mises' theory causes no difficulties; but the problem of application has still to be discussed.

The applicability of a model requires that the model has its counterpart in reality. This implies that there are two relevant aspects for the connection between model and reality. Firstly, we need input for the model, that is, certain theoretical variables should be given values derived from reality; and secondly, conclusions derived from the model should be translated back to real-world statements. In Von Mises' probability model this means that, from a finite number of observations, we should be able to derive the probability values in the collective, and from the probability values in new collectives (derived by means of the four elementary operations) we should be able to derive properties of finite samples. We shall first discuss the last part of this statement.

Assume that, according to our model, the probability of an event A is equal to p. Can we conclude anything about the relative frequency of occurrence r_n of event A in a finite sequence of n elements from this statement? The theorem that seems relevant in answering this question is the first law of large numbers. According to this law we have

$$\lim_{n\to\infty} p[|p - r_n| \leq \varepsilon] = 1$$

where ε is an arbitrarily small positive number. In the frequency interpretation, the theorem should be read in the following way: If an experiment is repeated n times (n large), we will find a relative frequency r_n; if this is done not once, but many times, then the value of r_n will appear to be close to p in almost all cases. Now, given the value of p, can we make any verifiable statement about the occurrence of a relative frequency r_n for n-times repeated trials? Most times r_n will be close to p, but for the particular sequence we are interested in, we simply do no know! Logically, it is impossible to construct a verifiable statement; every value of r_n is admitted by the theory. In practice, if n is large enough, we shall be convinced that the improbable event that r_n deviates from p substantially does not apply to our particular sequence. In other words, frequentists implicitly use the following rule: An event with a very low probability of occurrence will not occur at an individual trial.

Let us now discuss the qestion as to whether we are able to derive probability values in the collective from a finite number of observations. The theorem that seems relevant in this case is the second law of large number or Bayes's theorem. It runs as follows: if the observation of an n-times repeated alternative (e.g. coin-tosses) shows a relative frequency r_n of "successes" (heads) then, if n is sufficiently large, the chance that the probability of success p lies between $r_n - \varepsilon$ and $r_n + \varepsilon$ is arbitrarily close to one, no matter how small $\varepsilon (\varepsilon > 0)$. At first sight, this theorem very much resembles the first law of large numbers. It is, however, quite different. In the first law of large numbers, the probability is calculated that some relative frequency r_n lies in a fixed interval around a fixed value of p. In the second law, the probability is calculated that the value of p lies in an interval around an observed value r_n. Thus we are dealing with a probability distribution over p-values! Hence there must be a collective with the p-values as its elements!

The following example shows how such a collective with its corresponding probability distribution can be imagined. Consider an urn filled with a large number of coins. The coins are not all equal; some of them are biased others are not. Our inference problem can now be brought into the following frame work. A coin is taken from the urn and tossed n-times. Let the relative frequency of heads be equal to n_1/n. What then is the probability that this coin has probability p of falling heads? This probability is of course a probability in an appropriate collective which can be derived in two steps. Firstly, we construct the collective representing the drawing of a coin from the urn, tossing it n times, and determining the relative frequency r_n of heads.

This collective is two-dimensional because each element in it consists of a pair of numbers namely p_i and r_n. Secondly, the required collective is derived from this one by means of a partition. That is, the required collective is formed by taking only those elements of the original collective that have a second coordinate equal to $r_n = n_1/n$. The probability in which we are interested is then the frequency of pairs with a first coordinate equal to p in this derived collective. In order to solve our problem numerically, we must know the composition of the urn. We need the probability distribution in that collective which is the theoretical counterpart of the drawing of a coin, noting its p-value and putting it back in the urn. In this collective, each p value possesses its own probability. Here we see clearly how a probability distribution over p-values can be imagined. This distribution, which is often called the prior distribution of p, is of importance because the required probability depends on it. Unfortunately, however, this distribution is usually unknown and we seem to be in trouble. But we are saved by the remarkable fact that the influence of the prior distribution decreases with n and finally becomes negligible. This is the fundamental reason why inference for large n becomes possible without knowing the prior distribution exactly. It is also the reason why in our original statement of Bayes's theorem nothing is said about this prior distribution.

It is clear why Bayes's theorem seems to solve our problem because, roughly speaking, it says: if in one set of observations, n being a large number, the relative frequency of success (heads) is equal to r_n, then we expect with great certainty that p is equal to r_n. Again, nothing can be concluded logically on the basis of a specific element out of the collective $\{p_i, r_n\}$. For a specific element, the theory admits the possibility that p_i deviates substantially from r_n. An inference can be made, however, if we accept the rule previously introduced: an event with a very low probability will not occur at an individual trial. We should realize that another problem exists: if no prior distribution is known, how large should n be for the probability to be really low? It is an extra assumption that, if we apply this rule in a particular case, n is large enough.

When we shall be dealing with the subjectivistic approach towards probability, it will appear that the prior distribution plays a dominant role. It would not be surprising if many subjectivists think that the very assumption of a prior distribution is the thing that separates the subjectivistic school from the frequentistic approach. As far as Von Mises is concerned, this presumption is false. Von Mises explicitly makes the assumption of a prior distribution. The difference between Von Mises and the subjectivists is rather in the interpretation of the prior distribution if the prior distribution is not known before hand. Von Mises assumes a hypothetical collection of p-values. Inference procedures are only possible for so large a number of observations that the specification of the prior distribution is irrelevant. Subjectivists, on the other hand, use prior distributions to describe subjective knowledge about the true value of p or, alternatively, as a mathematical tool that is equivalent with the chosen way of learning from past experience. In principle, it is possible to specify the prior distribution, and inferences are no longer restricted to large n.

We now arrive at the following interesting conclusion. Probability statements about single observations are excluded by Von Mises. This implies that it is meaningless to say that the probability of a coin falling heads in the next throw is equal to 0.5 or to speak of the probability of a certain individual dying next year. <u>This fact is one of the most important consequences of his definition of probability as being a limit of a frequency in a collective</u>. This fact causes no difficulties as long as we are interested in abstract theory. The picture changes, however, as soon as we want to apply the theory to large though finite sequences because then we have to say something about a single finite sequence. <u>But any statement about some finite sequence is a statement about a single observation in a new compound collective</u>! Notice that the rule "an event with a very low probability of occurrence will not occur at an individual trial" is a statement about a single observation. Hence, by introducing this rule we leave the world of relative frequencies!

In the new world, statements about single events are possible given a certain amount of evidence. We can now ask ourselves what kind of rules we can expect to hold in this new world. Consider a number of events which, according to our rule, are thought to be impossible; then the disjunction (union) of these events is not necessarily impossible according to the same rule. This situation can be compared with a lottery: if the number of winning tickets is small in comparison with the total number of tickets (low relative frequency), it is thought to be impossible for any ticket to be a winning one. Nevertheless there must be some winning ticket! Apparently, the inference rule, as implicitly used by Von Mises, is only permissible if the addition rule for "impossible" events is not needed for the problem at hand. Actually, for mass phenomena any problem that has to do with frequencies can be transformed in such a way that the additivity rule is of no practical importance! As long as we restrict the domain in which the theory should be applied to mass phenomena where experience has taught us that rapid convergence exists (that is, the frequency limits are approached fairly rapidly), the inference rule causes no troubles and can therefore be used.

Hence we can conclude that, if we add a simple rule, Von Mises reaches his goal, namely a scientific theory of probability for mass phenomena. However, as economists, we are nor completely satisfied with the theory. Its domain of application is too restricted for this branch of science. Von Mises explixitly excludes probability statements about unique events, the outcome of one specific coin toss for instance. In many cases, the economist deals with precisely this type of problem such as in investment selection problems. Even in time-series analysis, we run into difficulties because probability statements only make sense if we are able to define a collective. In order to discuss the problem, consider the general linear model[1] with the following set of assumptions about the vector of disturbances u:

$$Eu = 0$$

$$Euu' = \sigma^2 I$$

These assumptions imply that we deal with a collective of disturbances that corresponds to a continuous label space. Every year, the disturbance takes a specific value. This can be compared with specific outcomes of tossings of a coin. If the collective of disturbances really exists, as has been

----

1) In matrix notation: $y = X\beta + u$.

assumed so far, we can construct for instance a 99% confidence interval for the prediction $\hat{y}$. According to Von Mises' theory, we should interpret this procedure as follows. If the disturbances were to be measured year after year, then in 99% of the cases, for given X and β, $\hat{y}$ will lie in this interval. With the aid of the inference rule introduced above, we can use this information for a decision for a specific year.

Objections against this procedure are two-fold. From the point of view of the frequentist, the problem at hand does not allow the application of the frequentistic theory: Neither the assumption of the existence of a collective, nor the assumption of probabilistic properties (e.g., distribution functions) can be verified if the number of observations is small, as in the case of time series analysis. Even conceptually, the number of observations can not be increased considerably, because the economic structure is changing in time. The procedure will lead to solutions consistent with the a-priori assumptions, but there is no way of knowing whether we are far from reality or not! From these considerations, it is clear that applications as in the one above are entirely in conflict with Von Mises' ideas.

From the point of view of the economist, Von Mises' theory is not sufficient to serve his purposes. A 99% confidence interval will often lead to such a wide range for the prediction $\hat{y}$ that it has no practical meaning. If the level of confidence would be chosen much lower, the confidence interval can no longer be interpreted, because we can no longer use the rule "the very improbable is impossible".

4. The Subjectivistic Approach; Fair Bets

Economics is often concerned with problems that demand a certain course of action. Many of the models that are used are introduced with the more or less implicit purpose of solving practical decision problems. Statistical procedures are used if a decision problem is burdened with uncertainty. If we are concerned with some mass phenomenon, we might well apply the frequentistic concept of probability: A life insurance company will calculate the premium on the basis of a frequentistic model of death rates. In many cases, however, the economist deals with problems that do not entirely have the character of a mass phenomenon, although a certain kind of repetition can be considered. If, for instance, the sales of some product are known for the last ten years, what can we conclude about the sales next year? Evidently, there is uncertainty involved, but we cannot properly speak of a mass phenomenon. In such a case we could of course try to define some "hypothetical" collective, but then we run into the problems mentioned in section 3. There are, however, situations where repetitiveness is completely absent. To give an example, consider the problem faced by a man who is contracted to send goods to the Netherlands at the moment that a harbour strike at the port of Rotterdam is threatening. If no strike occurs, he has to pay 10 units for shipment costs; if there is a strike at the moment that the goods arrive, he will lose 100 units; if the goods have to be shipped to another port, he has to pay 20 units for transportation costs. Evidently, the decision that is taken depends upon the "probability" that a strike will occur: if the harbour strike is thought to be improbable then the goods will be shipped to Rotterdam, and if the harbour strike seems to be nearly certain then another port will be used. The problem can be described in the following way: relative to the shipment of goods to an alternative destination, the shipment to Rotterdam presents a risk of losing 80 units if the harbour strike is actually announced, versus a possible gain of 10 units if there is no strike.
In this form, the decision problem may be considered as a bet that is made by the decision maker. If he accepts the bet, that is, if he actually sends the goods directly to Rotterdam, then he evaluates the situation in such a way, that he believes it to be worth the risk. This evaluation will be made on subjective grounds. The "probability" of a harbour strike is the man's subjective degree of belief that a harbour strike will actually occur.

The theory of subjective probabilities deals with decision problems and claims to be useful for all three cases mentioned above. The construction of a theory, however, is only interesting when an isolated decision problem is considered as one of a number of decision problems. It is important to define this reference class explicitly, because the criteria which should hold for a decision problem are derived from this class[1].

The transition from a realistic decision problem, like the shipment of goods when a harbour strike treatens, to a betting model about the harbour strike does not leave the reference class of possible decisions unaffected. In the betting model bets on an event E are considered simultaneously with complementary bets on E^c; it is possible to bet in such a way that one wins if E occurs and it is also possible to bet in such a way that one wins if E^c occurs. In our example, the complementary bet would correspond to a decision that pays off if there is a harbour strike, while there is a loss if no strike occurs. Such a decision does not exist in reality. Hence the reference class for the betting model is an <u>extension</u> of the reference class for the original problem. This fact leaves the possibility that the consistency requirements derived for the extended reference class may be too strong for the original reference class. Conceptually, the betting model is useful for the derivation of (subjective) probabilities; in practical situations people might deviate from this model because they think the betting model inappropriate.

Let us now discuss the betting model which has been developed by B. de Finetti [5] (in 1937) in more detail. We shall first pay some attention to the notation. The postulates, definitions and theorems will be numbered with the addition "F" to distinguish them from other approaches. The set of possible states of nature is denoted by {s}. The union of all s is the certain event S. Subsets of S are the events A, B, ..., E, ... The

----

1) It is noteworthy that this extension of an isolated decision problem to a class of decision problems is also met in the frequentistic probability theory: When, for instance, a hypothesis is tested in a frequentistic context, then the decision whether to reject the hypothesis or not is based upon considerations concerning a repetition of this decision under similar circumstances; the decision is not considered as an individual one (although in practice it might be unique) but as a specimen from a class of decisions.

complement of E with regard to S is written as E^c. A bet on E with stakes p:q is a bet such that q is gained if E occurs and p is lost if E^c occurs; this bet is denoted by $B_E(p:q)$.

Let $\mathcal{B}$ be the set of bets $B_E(p:q)$ for all $E \subset S$, p and q $\geq$ 0 but not simultaneously equal to 0. We shall take the following line of reasoning: First we construct subsets $\mathcal{B}_c$ of $\mathcal{B}$, containing consistent bets. The criterion for consistency is roughly that no combination of bets results in a net pay-off that is always unprofitable. Second, probabilities will be defined in terms of betting odds for bets in $\mathcal{B}_c$, and finally it will be proved that the ordinary rules of calculation for probabilities are valid.

Formally, we define subsets $\mathcal{B}_c$ by means of

D.F.1. $\mathcal{B}_c$ is a subset of the set $\mathcal{B}$ of all bets such that

(I) For any non-empty $E \subset S$ and for any value $p > 0$, there is some q such that $B_E(p:q) \in \mathcal{B}_c$. For any strict subset $E \subset S$ and for any value $q > 0$, there is some p such that $B_E(p:q) \in \mathcal{B}_c$

(II) If $B_E(p:q) \in \mathcal{B}_c$, then $B_{E^c}(q:p) \in \mathcal{B}_c$

(III) No finite number of bets in $\mathcal{B}_c$ on which we bet simultaneously may lead to a net result that is negative if E obtains ($E \neq \emptyset$), or that is negative or zero if E^c obtains.

The first part of this definition simply states that in the subset $\mathcal{B}_c$ bets are permitted on all events E, where the stakes can be chosen at an arbitrary level. The second requirement states that, if a person is willing to bet in accordance with $\mathcal{B}_c$, and if he is willing to bet on the occurence of E with the odds p:q, then he is also willing to accept the complementary bet on E^c. In such a case one speaks of a fair bet. It may be noted that fairness has nothing to do with objective fairness or intersubjective fairness; it is only the opinion of the person in question that matters. The set $\mathcal{B}_c$ is not the set of all bets that a rational person might accept, because it is very rational to accept unfair bets, as long as the advantage is on your side. The set $\mathcal{B}_c$ should rather be considered as a set of fair bets, in which inconsistent betting behaviour is excluded. The latter is achieved by requirement (III). There are weaker and stronger versions of this requirement. The stronger form is given above; the weaker form only requires that combinations of bets should never lead with certainty to a negative net result. De Finetti calls this the requirement of coherence, whereas (III) reflects the requirement of strict coherence. As will be seen, the difference between coherence and strict coherence lies in the admissibility of non-empty sets E having the probability value 0.

Now the following theorems can be derived:

T.F.1. If $B_E(p:q) \in \mathcal{B}_c$ and $B_E(p:q') \in \mathcal{B}_c$, where $E \neq \emptyset$, then $q' = q$

The proof is simple. Because of (II), $B_{E^c}(q':p) \in \mathcal{B}_c$. From the combination $B_E(p:q)$ and $B_{E^c}(q':p)$ it appears that the net result is equal to $q - q'$ if E obtains, and the net result is equal to $p - p = 0$ if E^c obtains, so by (III) we find $q - q' \geq 0$. In the same way, the combination $B_E(p:q')$ and $B_{E^c}(q:p)$ gives $q' - q \geq 0$, thus $q' = q$.

T.F.2. If $B_E(p:q) \in \mathcal{B}_c$, then $B_E(cp:cq) \in \mathcal{B}_c$ for any $c > 0$.

To prove T.F.2., we first assume that c is a rational number. In that case, c can be written as the quotient of two integers, say, $c = c_1/c_2$. From (I) it appears that there is some $B_E(cp:x)$ in $\mathcal{B}_c$, and we have to prove that $x = cq$. Now consider a combination of c_2 identical bets $B_E(cp:x)$, together with c_1 identical bets $B_{E^c}(q:p)$. If E obtains, the net result is $c_2x - c_1q$; if E^c obtains, the net result is $c_1p - c_2cp = 0$. Therefore, $c_2x - c_1q \geq 0$. In the same way, a combination of c_2 identical bets $B_{E^c}(x:cp)$, together with c_1 identical bets $B_E(p:q)$ gives $c_2x - c_1q \leq 0$, and therefore $x = cq$. Suppose next that c is not rational. Then there are rational numbers c' and c'' such that $c' < c < c''$, where the difference $c'' - c'$ can be chosen arbitrarily small. We have proved already that $B_E(c'p:c'q)$ and $B_E(c''p:c''q)$ are in $\mathcal{B}_c$. Again, let $B_E(cp:x)$ denote the bet on E that belongs to $\mathcal{B}_c$. From the combination of the bets $B_E(cp:x)$ and $B_{E^c}(c'q:c'p)$ the net result appears to be $x - c'q$ if E obtains, and $c'p - cp$ if E^c obtains. The latter value is negative, so $x - c'q > 0$ must hold, according to (III). In the same way, the combination of the bets $B_E(c''p:c''q)$ and $B_{E^c}(x:cp)$ leads to the finding that $c''q - x > 0$. Taking the two results together, we find $c'q < x < c''q$, and because c' and c'' can be chosen arbitrarily close to c, we have $x = cq$.

Subjective probabilities can now be defined upon the class $\mathcal{B}_c$:

D.F.2. If $B_E(p:q) \in \mathcal{B}_c$, then $P(E) = p/(p+q)$

From D.F.1.(I) it appears that there is a probability value for any E. From theorems T.F.1. and T.F.2. it appears that the probability value is defined in a unique way. Probabilities, defined in this way, fulfil the ordinary rules of calculation for probabilities (up to finite additivity):

T.F.3. (i) $0 \leq P(E) \leq 1$ for any $E \subset S$
(ii) $P(S) = 1$
(iii) if $A \cap E = \emptyset$, then $P(A \cup E) = P(A) + P(E)$

The first statement follows directly from the definitions (stakes have been assumed to be positive). The second statement is equivalent to $B_S(p:q) \in \mathcal{B}_c$ only if $q = 0$. In terms of the complementary bet, this implies $B_\emptyset(q:p) \in \mathcal{B}_c$ only if $q = 0$. Now suppose that $q > 0$ holds, then this bet would result in a certain loss of the amount q, which contradicts D.F.1. (III).
As to the third statement, let $D = A \cup E$. Consider the bets $B_A(p_A:q_A)$, $B_E(p_E:q_E)$ and $B_{D^c}(q_D:p_D)$ in $\mathcal{B}_c$. According to T.F.2. modifications of these bets can always be made such that the total stakes are equal, so we may assume without loss of generality that $p_A+q_A = p_E+q_E = p_D+q_D = M$. Now if A prevails, the total result can be written as $+q_A-p_E-q_D$, and using the equality of the total stakes this equals $-p_A-p_E+p_D$. The same result is obtained if E prevails. If $(A \cup E)^c$ prevails, the total result is $-p_A-p_E+p_D$, so this result is obtained with certainty. From D.F.1. (III) it appears that $-p_A-p_E+p_D \geq 0$. From the complementary bets on A^c, E^c and $(A \cup E)$ we obtain likewise the inequality $+p_A+p_E-p_D \geq 0$. But then, $p_D = p_A+p_E$. For any bet, the sum of the stakes equals M, so we have $P(A) = \dfrac{p_A}{p_A+q_A} = \dfrac{p_A}{M}$, and $p_A = M.P(A)$; in the same way, we have $p_E = M.P(E)$ and $p_D = M.P(D)$. It may be concluded that $P(D) = P(A) + P(E)$, which was to be proved.

As a direct consequence of the requirement of strict coherence, we find[2]:

T.F.4. For any $E \subset S$, $E \neq \emptyset$, $P(E) > 0$ must hold.

Suppose that this theorem is not true, then there must be some non-empty $E \subset S$ with $P(E) = 0$. But then, there is a bet $B_E(p_E:q_E)$ in $\mathcal{B}_c$ with $p_E = 0$. Consider the complementary bet $B_{E^c}(q_E:p_E)$. According to D.F.1. (II) this bet must also be an element of $\mathcal{B}_c$. If E prevails, this bet results in $-q_E$; if E^c prevails, this bet results in p_E, and this has the value 0. Now such a bet is excluded by D.F.1. (III) and cannot be an element of $\mathcal{B}_c$. A contradiction is found, so T.F.4. must be true.

From the assumed existence of the set $\mathcal{B}_c$ we have derived a finitely additive probability measure. Now we must turn to the question whether sets $\mathcal{B}_c$ do in fact exist. There are two sides to this question. First it must be considered whether such sets can exist in a mathematical sense. As a second step, it must be considered whether such a set exists in an empirical sense. It is reasonable to assume that a set $\mathcal{B}_c$ actually reflects the betting behaviour of people, at least approximately and under suitable circumstances? As to the first part of the question, one may apply the results of measure theory. It is well-known that a finitely additive probability measure exists for the set of all subsets of S. (see for instance [8]). For any $E \subset S$, let p_E denote the measure of E. Consider the set of all bets $B_E(M_E p_E : M_E(1-p_E))$ with $E \subset S$ and $M_E > 0$. It is easily seen that this set satisfies the requirements I and II of D.F.1.. It remains to be shown that requirement III is satisfied as well. From measure theory it is known that an operator, called the <u>expectation</u>, can be defined for random variables X, Y, (for a definition see Savage App. 1), such that

(a) $\mathcal{E}(\alpha X + \beta Y) = \alpha \mathcal{E}(X) + \beta \mathcal{E}(Y)$ where α and β are real numbers

(b) $\mathcal{E}(\chi_E) = p_E$ where $\chi_E = 1$ for all $s \in E$
$\chi_E = 0$ for all $s \in E^c$

(c) $\mathcal{E}(X) \geq 0$ whenever $P(X(s) < 0) = 0$

----

2) If we assume coherence instead of <u>strict</u> coherence, T.F.3. can still be proved. Only T.F.4. is no longer valid.

Consider a bet of the form B_E ($M_E p_E$: M_E (1 - p_E)). If E prevails, this bet will result in a gain M_E (1 - p_E), and this gain can be considered as a random variable M_E (1 - p_E). χ_E where the random variable χ_E is defined above. If E does not prevail, the bet will result in a negative - $M_E p_E$, and this can be considered as a random variable - $M_E p_E \cdot \chi_{E^c}$. The net result is again a random variable, and using (a) and (b) it is easily seen that the expected value is equal to

$$(M_E(1-p_E)).p_E - (M_E p_E).(1-p_E) = 0.$$

The net result of any finite combination of bets can again be considered as a random variable R which is the sum of random variables R_i for the individual bets i. But then, evidently,

$$\mathcal{E}\ (R) = \mathcal{E}\ (\sum_i R_i) = \sum_i \mathcal{E}\ (R_i) = 0$$

Now if requirement (III) is violated, the net result of the combination of bets is negative if some event A prevails, and is negative or zero if A^c prevails. But then, evidently, the expectation of the net result is negative, and we have proved above that this expectation must be equal to 0. Therefore, the betting system as described above fulfils the requirements (I) through (III), and the definition of $\mathcal{B}_c$ is not logically empty.

As to the empirical significance of $\mathcal{B}_c$, we shall postpone our comments until the mathematical part of the theory has been extended with the concept of conditional probabilities. These could be defined in a technical way by putting $P(A/B) = P(A \cap B)/P(B)$ for $P(B) \neq 0$. Such a definition would be meaningless if no interpretation of it were given. Conditional probabilities are conceptually new, rather than that they are a technical derivative of the original probability concept. Essentially, we must give an interpretation of restricted state spaces, and we must give the rules that relate the original state space with the restricted state space. Instead of using the technical definition, the above considerations will be the basis of our line of thought; the formula $P(A/B) = P(A \cap B)/P(B)$ will be obtained as a result.

Let S be the original state space and let $A \subset S$ ($A \neq \emptyset$). Then we define

D.F.3. The bet $B_{E/A}(p:q)$, that results in a gain q if $E \cap A$ occurs, that results in a loss p if $E^c \cap A$ occurs, and that is called off if A^c occurs, is a bet on E, conditional on A.

It is noteworthy that the presumption that A occurs is hypothetical. As soon as A is actually observed, we might change our opinions and determine new values $P'(E \cap A)$ and $P'(E^c \cap A)$ on a state space that is restricted by factual observation of A. In section 6, when dealing with the use of information, we shall have to explicitly introduce the assumption that P(E/A) and $P'(E \cap A)$ are identical, in other words, that there is no change from the situation in which A is presumed to be true to the situation where A has actually been observed.

The relation between the original bets and the conditional bets is derived from the requirement that no finite combination of bets and conditional bets may result in a certain loss. Let $\tilde{B}$ be the set of all conditional bets $B_{E/A}(p:q)$ with p and q $\geq$ 0, p and q not simultaneously equal to 0, E and A in S and A not empty. For A identical with S, it appears that the class B is a subset of $\tilde{B}$, in other words, the original bets can be considered as bets that are conditional on S. A rational betting model can now be defined by the consideration of a subset $\tilde{B}_c \subset \tilde{B}$ fulfilling the following requirements:

D.F.4. $\tilde{B}_c$ is a subset of the set $\tilde{B}$ of all conditional bets such that

(I) For any non-empty $A \subset S$, for any $E \subset S$ with $E \cap A \neq \emptyset$ and for any value $p > 0$, there is some value q such that $B_{E/A}(p:q) \in \tilde{B}_c$.
For any non-empty $A \subset S$, for any subset $E \subset S$ with $A - E \neq \emptyset$, and for any value $q > 0$, there is some value p such that $B_{E/A}(p:q) \in \tilde{B}_c$.

(II) If $B_{E/A}(p:q) \in \tilde{B}_c$, then $B_{E^c/A}(q:p) \in \tilde{B}_c$.

(III) No finite combination of bets in $\tilde{B}_c$ may lead to a net result that is negative if $E \cap A$ obtains ($E \cap A \neq \emptyset$), and that is negative or zero if $E^c \cap A$ obtains.

Theorems T.F.1. and T.F.2. remain valid when bets $B_E(p:q)$ are substituted by bets $B_{E/A}(p:q)$. Conditional probabilities can now be defined directly from conditional betting situations, and the definition is unique:

D.F.5. If $B_{E/A}(p:q) \in \tilde{B}_c$, then the conditional probability P(E/A) =

$= p/(p+q)$.

The requirement D.F.4. (III) leads to the well-known formula for conditional probabilities.

T.F.5. $P(A).P(E/A) = P(A \cap E)$

For the proof of this theorem, consider the three following bets: the bet on E, conditional on A, with stakes $M.P(E/A):M.(1-P(E/A))$; the bet on $(A \cap E)^c$ with stakes $M.(1-P(A \cap E)):M.P(A \cap E)$; and the bet on A with stakes $M.P(E/A).P(A):M.P(E/A).(1-P(A))$. In the third case, the factor $P(E/A)$ is just some multiplication factor, as permitted by the analogy of T.F.2. For $P(E/A) \neq 0$ this is always possible. Consideration of the events $A \cap E$, $A \cap E^c$, $A^c \cap E$ and $A^c \cap E^c$ leads to the recognition that the net result of these three bets equals $M.(P(A \cap E)-P(E/A).P(A))$ for any of these four events. The exclusion of a certain loss implies the inequality $P(A \cap E) \geq P(E/A).P(A)$. The combination of the three complementary bets shows that $P(A \cap E) \leq P(E/A).P(A)$, and T.F.5. follows. For the case where $P(E/A) = 0$, the procedure cannot be followed, because we would have a multiplication factor in the third bet that is equal to 0. We have to prove T.F.5. in a different way. It is sufficient to show that in this case $P(A \cap E) = 0$, because then T.F.5. is trivially fulfilled. Now, if $P(E/A) = 0$, there is a bet $B_{E/A}(p:q)$ in $\tilde{B}_c$ with $p = 0$. Consider the complementary bet $B_{E^c/A}(q:p)$. This bet will result in gain 0 if A^c prevails, because in that case the bet is called off. It will result in gain p if $E^c \cap A$ prevails, where $p = 0$, and it will result in loss q if $E \cap A$ prevails. Because p and q cannot be simultaneously equal to 0, requirement D.F.4. (III) shows that this situation can only occur if $E \cap A = \emptyset$, and with this the proof of T.F.5. is completed.

As to the existence of the class $\tilde{B}_c$, the line of thought is the same as for D.F.1. Let P be a finitely additive probability measure for all subsets of S; if $P(A) \neq 0$ then conditional probabilities $P(E/A)$ can be derived from formula T.F.5. Consider a system of bets where the stakes are proportional to the conditional probabilities. Such a system fulfils the requirements I and II of D.F.4. For conditional bets $B_{E/A}(M.P(E/A)):M.(1-P(E/A))$ the expected result is equal to

$$P(E \cap A).M(1-P(E/A)) + P(E^c \cap A).(-MP(E/A)) + P(A^c).0 =$$

$$= P(E \cap A)M - P(E \cap A)M \frac{P(E \cap A)}{P(A)} + (P(A) - P(E \cap A)).(-M \frac{P(E \cap A)}{P(A)}) = 0$$

But then, any finite combination of conditional bets leads to a net result with expected value 0. This prevents any finite combination of conditional bets from resulting in a certain loss, as defined in D.F.4. (III).

It has now been proved that the betting model is not logically empty. It remains to be seen whether the betting model has an empirical meaning. Probabilities are not uniquely defined by D.F.4. Uniqueness is only obtained by the assumption that people actually assess probabilities conform the requirements of D.F.4. Therefore, we state explicitly:

Ass. F.1. Under "suitable" circumstances, people can be brought to a betting behaviour that is reflected more or less accurately in the normative system of D.F.4.

It has already been pointed out that a theory of betting is not necessarily equivalent to a theory of decision-making in real-life situations. The set of possible decisions may influence the theory that is needed. Moreover, there is a psychological problem: If the "suitable circumstances", as mentioned in Ass. F.1., are widely different from the circumstances in real-life situations, then it might be possible that the evaluation of subjective probabilities differ in the two cases. Comments on these points are centred around the following assumption:

Ass. F.2. The normative system, as described in D.F.4., and measured under "suitable circumstances", as assumed under Ass. F.1., is also valid for real-life situations.

The first assumption is essentially concerned with problems of measurement; the second assumption points to the range of possible applications. There are different aspects of the measurement problem:

a) People might be unwilling to play the game. A hardheaded frequentist would perhaps refuse to spend his time on such a silly thing as the measurement of his subjective feelings, the results of which might easily be "misused" in his opinion. Here the "suitable circumstances" might offer a way out. People might be interested in psychological experiments, and this could give them the motivation to play along; other people might perhaps be persuaded when money is offered to them as a reward for their participation.

b) Even if people are cooperative, they might be vague in the assessment of probabilities. This can be a serious difficulty when real-life propositions are concerned. The probability of the occurrence of a harbour strike might be for some person "something between 10 and 30 per cent", and he happily admits that his personal assessment of the probability of the non-occurrence is "something between 90 and 70 per cent". In other words, there is a range of admissible bets that are considered fair. In a rational betting model such behaviour is excluded; nevertheless, the person in question might be unable to determine a fair bet in a unique way. "Suitable circumstances" can be developed such that people are worse-off if they deviate from their "correct" personal probabilities. Rational people, in the idealized sense, would then react by giving the "correct" probability value. In reality, however, people might still be unaware of their "correct" personal probabilities.

c) In actual behaviour, the consistency requirements may be voilated. Even when the subject succeeds in the determination of bets that appear fair to him, then these bets must still be related with each other in such a way that no inconsistencies occur. Often, fair bets will be determined for a small number of events only; other probabilities are then determined by means of interpolation. The interpolated results must be checked as to whether they correspond to the subject's personal feelings. In this way one may reach a consistent system, if the bets that generate the system are chosen in such a way that they are intuitively understandable.

d) People will almost certainly deviate from the system in the case where stakes are really high. Suppose that someone thinks a bet $B_E(1:1)$ fair, where gain and loss are both one dollar. The same person will think such a bet to be unfair if the gain and loss are both $10,000 because the loss of such an amount of money is more important to him than the possible gain of such an amount of money. In other words, the multiplication rule T.F.2. is violated. For the measurement of subjective probabilities this is not necessarily dramatic, because we do not need the complete system for the derivation of subjective probabilities: We might measure them for the case where the amount of money involved is limited, and we might extend the field of application of these probabilities to cases where larger amounts of money are involved by introducing a utility function. But then, we run into the following problem:

e) If the stakes are low, people might easily determine fair bets which deviate from their real feelings. The effort of careful consideration is hardly worth while if the consequences hardly matter. When people are unable to determine fair bets at high stakes but accept fair bets at low stakes, we do not know whether the latter is done because it reflects their personal feelings correctly, or because the stakes are so low that careful consideration is not worth while.

The remarks made so far make it clear that the "suitable circumstances", as mentioned in assumption F.1., must be created with care. This has a disadvantage for the transition to real-life problems also:

f) the assessment of probabilities under laboratory conditions may be different from the assessment of probabilities within a real-life context. Probabilities are derived from an intuitive process in which different influences upon some real-life situation are weighed; when this process is carried out in circumstances that are remote from the situation in which the actual decision has to be taken, this process may be disturbed. It is by no means certain that a bet on the occurrence of a harbour strike will be similar to a decision that is taken when a harbour strike threatens even if the formal description of the two is the same.

It has been stated already that a unique probability measure in the fair betting model can only be defined on the basis of the measurement of people's behaviour. Therefore the problems mentioned above, which are met in the process of measurement, are relevant. They do not, however, give grounds for the complete rejection of the theory of personal probabilities. Problems like unwillingness, vagueness, carelessness in the evaluation and comparability with real-life decision problems might be overcome by educating people. Psychologists may find the "suitable circumstances" in which people respond in the required way. Some of the problems may be solved by the creation of circumstances that are more directly related to the real-life decision problems; if necessary, the betting model can then be exchanged for another model, and the postulates adapted. In the next section, we shall see that Savage has developed an axiomatic basis for subjective probabilities where fair bets do not play a part.

There are two other points that must be raised, which are more fundamental than problems of measurement.

g) In the example of a harbour strike, we have seen that sending the goods to the town where the strike may occur can be considered as the acceptance of a bet, whereas the complementary bet is void of any practical meaning. In the betting model, both bets and complementary bets play a part. Is the use of the betting model then appropriate for the harbour strike situation?

Apart from measurement problems, has the betting model (or modifications of it) general validity? Here we come to the point that has been raised by Kyburg [13]. He states that there are many situations in which our uncertainty cannot be expressed by betting quotients. Suppose that we are concerned with some long-term problem, for instance the exhaustion of the copper supplies of the world. Our certainty will then not be reflected in a betting game, because the pay-off will take place long after our death. Only hypothetical bets with a hypothetical pay-off will then be of any help. This argument is the more pressing if the truth of a general law is concerned: Such a law can only be falsified, and can never be verified. Therefore, the only reasonable bet is one with stakes 0 on the proposition that the theory is true; otherwise, we are betting money with the certainty that there will never be a pay-off. The other side of the bet is equally disadvantageous: if the law is not completely implausible it would be nonsense to let the stakes be positive, if the opponent offers no money at all. We are now in the degenerate situation of the bet $B_L(0:0)$, where L stands for the truth of a general law. Even in this situation one might try to think of hypothetical bets. Such reasoning, however, evades the fundamental problem: If we are faced with an uncertain situation, in which bets have no empirical meaning, there is no reason to require that the consistency requirements for bets should hold. Therefore, we state:

h) The betting model is only appropriate for at least in concept-verifiable statements.

Some of the problems raised above are - at least partially - solved by adapting the set of axioms. See, for instance, the later work of De Finetti (especially [6]), and the next section in which Savage's approach is treated. Yet, the heart of the matter remains unchanged: In real life decision problems, the set of possible acts is often too limited to obtain numerical probabilities. Therefore, the set of acts must be extended with artificial acts as, for instance, bets. But then we run into the problem that the behaviour of people under artificial circumstances may differ essentially from their behaviour under realistic circumstances.

Up till now, we have paid no attention to the most fundamental attack on the subjectivistic approach. Frequentists claim that their theory is objective, unburdened by personal taste. Therefore, their theory is real "science" whereas the subjectivistic approach is too personal to have scientific value. This discussion is one about the roots of science. At this stage, we cannot go into this discussion, because the frequentists deal with repetitiveness, whereas the subjectivists, as treated so far, deal with problems where repetitiveness is not an essential feature. It is only after dealing with the subjectivistic approach to phenomena that are (mainly) characterized by repetitiveness, as considered in the "mixed approach" (section 7) and in the normative approach (section 8), that this point can fruitfully be discussed.

5. Savage's Axiomatic System

L.J. Savage [19], derives numerical probabilities from a system of ordering relations. Savage's contribution is of special importance because of the context in which the problem is developed. He bases his system on a preference that we have for one act over another in an uncertain situation. It is assumed that such preferences will obey certain consistency rules. These rules are laid down in the postulates. The postulates can be interpreted in two ways. In the behavioralistic approach, it is assumed that the postulates reflect more or less the actual behavior of mankind in situations of uncertainty. Whether or not people react in a consistent way is a problem belonging to the field of psychology. We are more interested in the normative interpretation: Given some consistency rules, in what way should people react to a choice between acts in a situation of uncertainty? This is apparently a decision problem. Usually, the solution of decision problems is derived from probability theory. Savage goes the other way around. If it is possible to solve a decision problem, there must be a preference relation defined on the set of acts; if this preference obeys some consistency rules, then it can be shown under what other conditions numerical probabilities must exist. The ordering among acts implicitly defines the underlying notion of probability. Because the ordering among acts is subjective, the underlying probabilities are also subjective.

Let us consider Savage's definitions and postulates more closely. In the real world, or rather the relevant part of the real world, several possible states can be distinguished. For instance, when betting on one throw of a die, the possible states are the results "1" through "6". When confronted with a possible harbour strike, there are (roughly) two states: the strike will or will not occur. Uncertainty can be defined as a lack of knowledge about the state of the world that obtains. Now, any act will result, depending on the state of the world that obtains, in a specific consequence. Hence, we can define an act mathematically as a mapping of the set of states into the set of consequences.

The definition of an act is extended in the following way:

D.S.1. Any mapping of the set of states into the set of consequences is called an act.

If we make an even bet of one dollar on the outcome of a coin, there are two states of the world, the outcomes "heads" and "tails"; the two consequences are "winning a dollar" and "losing a dollar". According to the definition, there are four possible acts. Only two of these are realistic in a betting situation. If we put our dollar on "heads", state "heads" is mapped on the consequence "winning a dollar" and state "tails" is mapped on the consequence "losing a dollar". The second realistic act is to put our dollar on "tails". Moreover, there are two hypothetical acts, viz. winning a dollar whatever outcome results and losing a dollar whatever the outcome.

The first postulate asserts the existence of a preference ordering among <u>all</u> acts, either realistic or hypothetical.

P.S.1. In the class of acts, a complete ordering is defined, so if f and g are acts, then either $f \leq g$, or $g \leq f$, or both hold.

In our later comments, we shall consider more closely the implications of this assumption in a normative interpretation of the postulates. At this stage, we can restrict ourselves to the remark that a preference among realistic acts at least seems to be necessary for a consistent solution to a decision problem. As to the behavioralistic approach, it is rather doubtful whether the postulate would be fulfilled in complicated problems.

It is clear that the ordering among acts entails an ordering among consequences. For this purpose, identify each consequence c with the so-called constant act f_c that maps all states on the consequence c. The constant act f_c results with certainty in consequence c, and the ordering among the class of such constant acts can be identified with the ordering among consequences.

Other postulates will be needed to restrict further the class of possible orderings among acts. These postulates reflect common sense principles about preferences in a situation of uncertainty. If, for instance, the set of states S is partitioned into the subsets B and S-B and if the acts f and g are identical on the subset S-B, then the preference relation between f and g will be completely determined by the specification of f and g on the relevant part B of S. The same is true for any other pair of acts f' and g' when f' and g' are identical upon S-B. Apparently, the following postulate reflects this common sense principle:

P.S.2. If f is identical with g on S-B; if f' is identical with g' on S-B; if f and f' are identical on B and if g and g' are identical on B; if, moreover, $f \leq g$, then $f' \leq g'$ also.

This postulate makes it possible to define in a unique way an ordering relation among acts if the set S is restricted to the subset B.

D.S.2. $f \leq g$ given B; if and only if $f' \leq g'$, where f' and g' are modifications of f and g so that f and f' are identical on B, g and g' are identical on B, and f' and g' are identical on S-B.

From P.S.2. it appears that, if the ordering given B were to define with the help of another pair of modifications, say f" and g", under the same rules of construction, then the same ordering given B would result as in the case where f' and g' are used.

It can easily be checked that $f \leq g$ given B is a complete ordering. Apparently, only the specification of f and g on the subset B is relevant for the ordering relation $f \leq g$ given B. It should be noted that the ordering relation $f \leq g$ given B is interpreted as a (conditional) preference and not merely as a mathematical derivation.

So, if the preference ordering among acts, defined on S, is given and if the extra information that S-B does not obtain is available, the conditional preference is given by $f \leq g$ given B.

An analogon for events with probability measure 0 is found in the definition of a null set:

D.S.3. The set B is null if and only if for all acts f and g holds $f \leq g$ given B.

Apparently, if B is null, both $f \leq g$ given B and $g \leq f$ given B, so all acts are equivalent given B.

We have already seen that an ordering among consequences is obtained by considering the ordering among constant acts f_c which result with certainty into the consequences c. Now, if by some extra information the set S is reduced to the subset B, acts that are constant on B result with certainty into the corresponding consequences. This again would imply a (new) ordering among consequences. We are not sure, however, taht the original ordering among consequences is the same as the new one. This is asserted in postulate 3.

P.S.3. If f_c and $f_{c'}$ are constant acts; if B is not null; then $f_c \leq f_{c'}$ if and only if $f_c \leq f_{c'}$ given B.

The fourth postulate is trivial:

P.S.4. There is at least one pair of consequences c_1 and c_2 so that $c_1 > c_2$.

If all consequences were equivalent, no decision problem would be left; all acts would be equally attractive!

In set theory, subsets B of S can be described by their characteristic functions. A characteristic function has the value 1 for all points in B, and the value 0 for all points of S-B. In the same way, as soon as P.S.4. is satisfied, we can define "characteristic acts" f_B that map all points of B on c_1 and all points of S-B on c_2[1].

Apparently, a characteristic act f_B is constant on B and constant on the complement S-B. The class of characteristic acts, defined for the consequences $c_1 > c_2$, is completely ordered. Because any subset B corresponds to one and only one characteristic act from this class, this ordering can be transposed directly to the set of subsets of S. The ordering relation $A \geq B$ for the subsets A and B of S can be interpreted as a "more probable

----

[1] The terminology of characteristic acts is not used by Savage; here it is introduced to stress the similarity with the theory of sets.

than" relation. If there is a game that results in \$100 earnings if A obtains and \$100 losses if S-A obtains, if there is another game that results in \$100 earnings if B obtains and \$100 losses if S-B obtains, and if we prefer the first game to the second, then apparently A is considered more probable than B.

The postulates P.S.1. through P.S.4. do not guarantee that, if another pair of consequences $c_3 > c_4$ is chosen, the same ordering of subsets would be found. For this reason, a fifth postulate is introduced:

P.S.5. If the characteristic acts f_A and f_B are defined for the consequences $c_1 > c_2$, if the characteristic acts g_A and g_B are defined for the consequences $c_3 > c_4$ and if, moreover, $f_A \leq f_B$, then $g_A \leq g_B$.

Apparently, acts are assumed to be ordered in such a way that, if a \$100 bet on A is preferred to be bet on B, then a \$1000 bet on A must also be preferred.

The "more probable than" relation among subsets of events has the following three properties:

(i) the ordering is complete
(ii) if $B \cap D = C \cap D = \emptyset$ (is the empty set), $B \leq C$ if and only if $B \cup D \leq C \cup D$.
(iii) $B \geq \emptyset$; $S > \emptyset$.

The first property is a direct consequence of the definition of $A \leq B$. The second property is a direct consequence of P.S.2. (f_B and $f_{B \cup D}$ are identical on B; this implies that f_B and $f_{B \cup D}$ are identical on $B \cap C$. Because $C \cap D = \emptyset$, f_B and $f_{B \cup D}$ are also identical on $(S-B) \cap C$. Therefore, f_B and $f_{B \cup D}$ are identical on $B \cup C$. In the same way, f_C and $f_{C \cup D}$ are identical on $B \cup C$. Moreover, f_B and f_C are identical on $S-(B \cup C)$, and $f_{B \cup D}$ and $f_{C \cup D}$ are identical on $S-(B \cup C)$. The use of P.S.2. completes the proof).

The proof of the first part of (iii) is obvious as soon as it is understood that $f_\emptyset$ is a constant act on S. Postulate P.S.3. then leads to the required proposition. The proof of the second part of (iii) is a direct result of P.S.4.

The structure of the ordering among events is determined by properties (i) through (iii). From these properties, we can derive intuitively clear propositions such as

(a) if $B \subset C$, then $\emptyset \leq B \leq C \leq S$

(b) if $B \leq C$, and $C \cap D = \emptyset$, then $B \cup D \leq C \cup D$.

From this point, we will omit the details and give a general outline of the construction of numerical probabilities. First, suppose that S can be partitioned into n mutually exclusive subsets $S_1, \ldots, S_n$ that are all equivalent. Koopman [11,12] uses the term n-scales for such partitions. Evidently, we can give a probability value of $\frac{1}{n}$ to each S_i. We can always find a union of r elements of the partition that is infra-probable to some given event A, such that the union of r+1 elements is more probable than A. Under certain conditions, the limiting value of $\frac{r}{n}$ exists, and the numerical probability of A is defined as this limiting value. Instead of approximating P(A) by unions of elements with the exact numerical probabilities $\frac{1}{n}$, we could just as well use unions with numerical probabilities of approximately $\frac{1}{n}$, as long as the approximation error vanishes for large values of n. Such approximations are used in Savage's concept of almost uniform partitions:

D.S.4. An n-fold almost uniform partition of B is an n-fold partition such that the union of any r elements is not more probable than the union of any r+1 elements.

It can be proved that the numerical probability of a union of r elements of an n-fold almost uniform partition is approximately equal to $\frac{r}{n}$. The introduction of almost uniform partitions has two advantages. Firstly, the philosophical problem as to which cases we may expect equal probabilities in is evaded. It seems difficult to define events with equal probabilities outside the field of games. Secondly, the weaker assumption of the existence of almost uniform partitions gives more insight into the essentials of the procedure leading to numerical probabilities. We now quote the main theorems needed for the construction of numerical probabilities. Proofs are omitted; occasionally, intuitive explanations are given. For details the reader is referred to [19].

(a) If n-fold almost uniform partitions of S for arbitrarily large values of n exist, then an m-fold almost uniform partition for any positive integer m exists. Indeed, if S can be partitioned in, say, 1000 events that are approximately equally probable, then it is possible to combine them in such a way that, say, three groups result that are equally probable up to a small approximation error (in numerical probability this error would be of the order 1/1000). These three groups certainly form a 3-fold almost uniform partition. It may be noted that a similar proposition is untrue for n-scales with equally probable events.

(b) If n-fold almost uniform partitions of S for any positive integer n exist and if there were to be a numerical probability measure P such that $A \leq B$ implies $P(A) \leq P(B)$, then such a probability measure is unique.

Indeed, if a probability measure exists then the combination of r elements of an n-fold almost uniform partition will have a numerical probability of about $\frac{r}{n}$. In fact, this value can be shown to between (r-1)/n and (r+1)/n. Now let k_n/n be the largest number of elements such that the union C_k is infra-probable to some event B. Then $C_{k_n} \leq B < C_{k_n+1}$ and, by hypothesis, this implies $P(C_{k_n}) \leq P(B) \leq P(C_{k_n+1})$. Using the approximation error for C_{k_n} and C_{k_n+1} mentioned earlier, we find

$$\frac{k_n - 1}{n} \leq P(B) \leq \frac{k_n + 2}{n} \text{, or } P(B) - \frac{2}{n} \leq \frac{k_n}{n} \leq P(B) + \frac{1}{n}$$

If P(B) exists, then the limit of k_n/n exists, and P(B) is equal to this limiting value. So then P(B) is uniquely determined.

(c) If n-fold almost uniform partitions of S for any positive integer n exist, then a probability measure P such that $A \leq B$ implies $P(A) \leq P(B)$ exists.

The proof contains the following elements. Consider some specific sequence of n-fold almost uniform partitions (n = 1, 2, 3, ...). Evidently, for some given subset B of S, we will expect that $P(B) = \lim_{n\to\infty} \frac{k_n}{n}$ where k_n is again defined as the largest number of elements of the n-fold partition such that its combination is not more probable than B. Thus, we have to prove that

$\lim_{n\to\infty} \frac{k_n}{n}$ exists. Now consider the difference $\frac{k_m}{m} - \frac{k_n}{n}$. We want to prove that this difference vanishes for large values of m and n.

Apparently, it is necessary to relate the m- and n-fold partitions to each other to derive the desired result. A natural way to do so seems to be the consideration of the common refinement of the two partitions. However, generally, such a common refinement is not almost uniform and nothing can be concluded. As a way out, we can define new m- and n-fold almost uniform partitions so that the common refinement is indeed almost uniform. The proof is split up into two parts. For the new m- and n-fold almost uniform partitions, it will be proved that $\frac{k'_m}{m} - \frac{k'_n}{n}$ vanishes for large values of m and n; then it will be proved that $\frac{k_m}{m} - \frac{k_n}{n}$ vanish as well. The new m- and n-fold almost uniform partitions are constructed in a rather evident way: Consider an m.n-fold almost uniform partition and always take the unions of n and m elements of it respectively. It can easily be proved that the resulting m- and n-fold partitions are indeed almost uniform. For the new m-fold partition, we can define a new number k'_m, corresponding to nk'_m elements of the m.n-fold partition. The same can be done for the new n-fold partition. Apparently, $C_{nk'_m} \leq B < C_{n(k'_m+1)}$ and $C_{mk'_n} \leq B < C_{m(k'_n+1)}$ must hold, where C denotes unions of elements of the m.n-fold partition. From these two inequalities $C_{nk'_m} < C_{m(k'_n+1)}$ and $C_{mk'_n} < C_{n(k'_m+1)}$ follow. Because the m.n-fold partition is almost uniform, these inequalities imply $nk'_m \leq m(k'_n + 1)$ and $mk'_n \leq n(k'_m + 1)$. From this, it appears that $\frac{k'_m}{m} - \frac{k'_n}{n}$ vanishes for large values of m and n.

For the second part of the proof, unions of r elements of the original m-fold partition and of s elements of the newly constructed m-fold partition are denoted by C_r and C'_s respectively. The numbers k_m and k'_m are defined as before. Hence

$$C_{k_m} \leq B < C'_{k'_m+1}$$

in which $C'_{k'_m+1}$ is <u>any</u> union of $k'_m + 1$ elements of the new partition. As a special choice, let $C'_{k'_m+1}$ denote the union of the $(k'_m + 1)$ least probable elements of the (new) m-fold partition. Generally, it can be proved that the union of the s least probable elements of an m-fold partition is infra-probable to the union of the s most probable elements of another m-fold partition (this is true even for partitions that are not almost uniform). Consequently, we have

$$C_{k_m} \leq B < C_{k'_m+1}$$

where $C_{k'_m+1}$ denotes the union of the $(k'_m + 1)$ most probable elements of the original m-fold partition. Because the original m-fold partition is almost uniform, we can conclude that $k_m \leq k'_m + 1$; hence $\frac{k_m}{m} - \frac{k'_m}{m} \leq \frac{1}{m}$. In the same way, it can be proved that $\frac{k'_m}{m} - \frac{k_m}{m} \leq \frac{1}{m}$, and consequently $\frac{k_m}{m} - \frac{k'_m}{m}$, vanishes for large values of m. Apparently, it can be proved in the same way that $\frac{k_n}{n} - \frac{k'_n}{n}$ vanishes for large values of n, and this completes the proof of the existence of $\lim_{n\to\infty} \frac{k_n}{n}$. The limiting value is called P(B), and this function on the set of subsets B fulfills the requirements of a (finitely additive) probability measure:

(i) $P(B) \geq 0$ for every subset B

(ii) $P(S) = 1$

(iii) If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$

(i) and (ii) are evident from the definition of P(B). (iii) must be considered more closely. Given some n-fold almost uniform partition, we know that the combination of any $k_n(A) - 1$ of its elements is less probable than A. In the same way, any combination of $k_n(B) - 1$ of its elements is less probable than B. (Indeed, we know that there is some combination of $k_n(A)$ of its elements that is infra-probable to A, so all combinations of $k_n(A) - 1$ of its elements are less probable than A). But then any combination of $k_n(A) + k_n(B) - 2$ of (mutually disjoint) elements will be less probable than $A \cup B$. On the other hand, $A \cup B$ is less probable than any combination of $k_n(A \cup B) + 1$ of its elements. Hence

$$\frac{k_n(A) + k_n(B) - 2}{n} \leq \frac{k_n(A \cup B) + 1}{n}$$

and, for $n \to \infty$, it appears that

$$P(A) + P(B) \leq P(A \cup B)$$

On the other hand, we know that any combination of $k_n(A) + 1$ elements is more probable than A; any combination of $k_n(B) + 1$ elements is more probable than B. Hence, any combination of $k_n(A) + k_n(B) + 2$ (mutually disjoint) elements is more probable than $A \cup B$. Moreover, there is a combination of $k_n(A \cup B)$ elements that is infra-probable to $A \cup B$. From this, it appears that $P(A \cup B) \leq P(A) + P(B)$. With this, (iii) has been proved.

The relation between the quantitative probability measure and the qualitative probability ordering is expressed by

(iv) If $A \leq B$, then $P(A) \leq P(B)$.

Indeed, $C_{k_n(A)} \leq A \leq B < C_{k_n(B)+1}$, so $\frac{k_n(A)}{n} \leq \frac{k_n(B) + 1}{n}$ and for the limit, $P(A) \leq P(B)$.

It can be concluded that if n-fold almost uniform partitions for arbitrarily large values of n exist, then a unique, finitely additive probability measure P can be derived. However, it is not easy to verify the existence of such n-fold almost uniform partitions. Savage aims to find a necessary and sufficient condition for the existence of such partitions. Instead of postulating directly the existence of n-fold almost uniform partitions for arbitrarily large values of n, Savage introduces the following concept:

D.S.5. An ordering among events is <u>fine</u> if, for any $C > \phi$, a partition $\{E_i\}$ of S exists such that $E_i < C$ for all E_i[2].

It can be proved that, if an ordering is fine, then n-fold almost uniform partitions do exist for arbitrarily large values of n; consequently, the finitely additive probability measure P(A) can be defined. It may be noted, however, that $P(A) \leq P(B)$ does not necessarily imply $A \leq B$. In other words, it is possible that both $A < B$ and $P(A) = P(B)$ hold. In such a situation,

----

[2] From the original text, one is inclined to think that $E_i \leq C$ is meant; this is apparently an inaccuracy.

A should be very close to B, because if also $A \cup E < B$ were to hold for some $E > \phi$ (and $A \cap E = \phi$), we would have $P(A) < P(A) + P(E) = P(A \cup E) \leq P(B)$, in contradiction with the assumption[3]. Apparently, it makes sense to define

D.S.6. If $A \leq B$ and if there is no $E > \phi$, $E \cap A = \phi$, such that $A \cup E \leq B$, then A and B are almost equivalent.

D.S.7. An ordering is tight if each pair of almost equivalent events is equivalent.

It can be demonstrated that there are orderings that are fine but not tight. One could think for instance of an irrational preference for "lucky numbers". Perhaps we prefer to buy a lottery ticket with a number ending on 7 if there are no other reasons for determining our choice, whereas any "real" improvement in our probability of winning is preferable to this kind of superstition. The following postulate asserts that the ordering among events is both fine and tight:

P.S.6. If $B < C$, a partition $\{E_i\}$ of S exists such that $B \cup E_i < C$ for all E_i ($E_i > \phi$).

Apparently the ordering is fine if this postulate is fulfilled because for $B = \phi$ we are back in definition D.S.5. Moreover, the ordering is tight because the situation $A < B$ with $P(A) = P(B)$ is impossible: from $B \cup E_i < C$ it appears that $P(B \cup E_i) \leq P(C)$; because $E > \phi$ and therefore $P[E] > 0$, it follows that $P(B) < P(C)$.

The postulates P.S.1. through P.S.6. define a unique finitely additive probability measure. Conditional probabilities could be defined in terms of this probability measure in a classical way. However, up till now postulates and definitions have been based on an ordering relation among acts that reflects a personal preference. The definition of conditional probabilities in the form $P(A|B) = \frac{P(A \cap B)}{P(B)}$ would be void of any meaning

----

[3] This reasoning essentially depends on the fact that, if the ordering is fine, $E > \phi$ implies $P(E) > 0$.

if it could not be interpreted in terms of a personal preference. Therefore, it is better to derive the properties of conditional probabilities from such ordering relations. An ordering among acts given B has already been defined. So the characteristic acts f_A given B are ordered as well. It is easily verified that $f_A \leq f_C$ given B, if and only if $f_{A \cap B} \leq f_{C \cap B}$. But the latter inequality holds, if and only if $A \cap B \leq C \cap B$. So we can define

D.S. 8. $A \leq C$ given B, if and only if $f_A \leq f_C$ given B.

It appears that $A \leq C$ given B, if and only if $A \cap B \leq C \cap B$.
If, for the original ordering, the postulates P.S.1. through P.S.5. are valid, it is easy to prove that the ordering among events given B satisfy the following requirements:

(i) The ordering $A \leq C$ given B is complete (because the events of the form $A \cap B$ and $C \cap B$ are completely ordered).

(ii) If $D \cap A = D \cap C = \emptyset$, then $A \leq C$ given B, if and only if $A \cup D \leq C \cup D$ given B. (Indeed, $A \cap B \leq C \cap B$ if and only if $(A \cup D) \cap B \leq (C \cup D) \cap B$).

(iii) $A \geq \emptyset$ given B and $S > \emptyset$ given B.

The latter part of (iii) is only true if $S \cap B > \emptyset$; for that reason, it is assumed in the following that $B > \emptyset$. Apparently, the ordering $A \leq C$ given B fulfills the requirements of a qualitative probability. Moreover, it can be proved that

(iv) if P.S.6. is valid for the original ordering, then the ordering among events given B satisfies this postulate as well.

Indeed, if $A < C$ given B, then $A \cap B < C \cap B$. Application of P.S.6. reveals that there is a partition $\{E_i\}$ of S, such that $(A \cap B) \cup E_i < C \cap B$ for all E_i. But then certainly $(A \cap B) \cup (E_i \cap B) < C \cap B$. The left member can be written as $(A \cup E_i) \cap B$, so there is a partition $\{E_i\}$ such that if $A < C$ given B, then $A \cup E_i < C$ given B for all E_i, and this is exactly the contents of P.S.6.

From (i) through (iv), it appears that a unique, finitely additive conditional probability measure exists. This is denoted by $P(A|B)$. It can easily be verified that $\frac{P(A \cap B)}{P(B)}$ fulfills the requirements for this probability measure. Because of its uniqueness, we can write $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

At this stage, the definition of the probability concept in Savage's approach is complete. Let us reconsider the system as a whole. The first question to be answered is, as to whether the postulates should be interpreted in a normative way or in a behaviouralistic way. Savage admits that his postulates could be only fulfilled by an "ideal person".

In practice, the requirements of consistency are too heavy to bear for any real life person. Most acts are hypothetical, and therefore inaccessible for intuition. Moreover, acts defined on complicated subsets B of the universal event S will be void of any intuitional meaning. Savage remarks:

> "The main use I would make of the postulates is normative, to police my own decisions for consistency and, where possible, to make complicated decisions depend upon simpler ones."

This implies that the postulates should not be considered as a simple reflection of intuitive preferences, but rather as a system in which the intuitive preferences are imbedded. One could say that the intuitive preferences (perhaps a bit polished for consistency reasons) must generate the more extended system as given by the postulates.

With the aid of a simple example, we shall try to demonstrate that the necessary amount of intuition even in quite common situations is lacking. We only can find a way out by reformulating the problem. Suppose we want to derive a (qualitative) subjective probability for the occurrence of a harbour strike at Rotterdam. Two possible realistic acts are defined, we can ship some party of goods to another harbour or not.

Consequences are (1) loss of money because the ship is sent to another port; (2) delay because the goods are stranded in Rotterdam; (3) no loss of money and no delay. If the events "strike" and "no strike" are denoted by A and S-A respectively (S being the universal event), then we may conclude:
If the ship is sent to Rotterdam, this act maps A on (2) and S-A on (3); if the ship is sent to another port, this act maps S on (1). To derive qualitative probabilities, it is necessary to introduce hypothetical acts, such as the one that maps A on (3) and S-A on (2). This would imply no loss of money and no delay if a strike would occur and delay if there is no strike. This hypothetical act is void of any intuitive meaning, and it is impossible to compare it with a realistic act like sending the ship

to Rotterdam. The decision problem as defined above does not lead to qualitative probabilities.

The only way out seems to be the introduction of new consequences, like for instance the loss and gain of an amount of money. Such an extension gives the opportunity to use hypothetical acts that have an intuitive meaning. We may compare the act that results in a loss of $100 if a strike would occur and a gain of $100 if no strike would occur, with the act that results in a gain of $100 if a strike would occur and a loss of $100 if no strike would occur. But then, qualitative probabilities are defined on comparisons of hypothetical betting situations, and the original decision problem does not play a role anymore in the determination of qualitative probabilities. In other words, the determination of qualitative probabilities is independent of the original specific decision problem. Apparently, the simultaneous treatment of decision problems and the determination of qualitative probabilities is rather artificial.

Let us now face the problem, in what circumstances numerical probabilities can be derived: Either n-scales or (more generally) almost uniform partitions are assumed to exist. It is of importance, to discuss the implications of such an assumption. Suppose that we are interested in the distribution of the lenghts of recrutes. This is a much quoted example in the frequency approach. Apparently, it is very difficult to define events that are equally probable, unless we have information about frequencies in a long series of observations. But then, n-scales are not primitive notions; they result from observations, and the corresponding probabilities should have been defined in terms of these frequencies. In this example, n-scales are a result of numerical probabilities, rather than vice versa. As a pure subjectivistic example with a continuous sample space we can use the size of the market for a newly developed product. The marketing manager may have certain ideas about the market size, but it will be very difficult to define intervals such that to these intervals equal degrees of belief can be given. Perhaps this is possible for 2, 3 or 4 intervals, but quite clearly the intuition of the marketing manager will be much too vague to define equally probable intervals that are arbitrarily small.

Savage suggests the following way out. We can extend the universal event with the outcomes of large sequences of coin tosses. Under the assumption that the postulates hold even for this extended class of events, numerical probabilities can be derived. The implication of this assumption is that real events can be compared as to their probability with a "canonical experiment". By introducing this concept we reach the ideas of Pratt, Raiffa and Schlaifer which we shall briefly discuss in the next paragraph.

6. Reference Lotteries: The Approach of Pratt, Raiffa and Schlaifer

When dealing with Savage's theory, we have omitted all references to the concept of utility. This could be done very easily, because probabilities are defined by Savage independent of utilities. In the axiomatic system as given in the paper of Pratt, Raiffa and Schlaifer [17], the concepts of utility and probability are defined simultaneously. We shall restrict ourselves here to those parts, that are relevant for a definition of the probability concept.

The aim of Pratt, Raiffa and Schlaifer is to develop a method to obtain a consistent subjective probability model for real life situations. Essentially, the theory boils down to a comparison of real-life events with a hypothetical canonical experiment. It is supposed that one can imagine an experiment characterized by a rectangular distribution on the interval $[0,1]$. Given two consequences c^* and c_*, we can consider the class of lotteries that result in c^* for the interval $[0,\alpha)$ and in c_* for the interval $[\alpha,1]$. For real-life events E_0 one can define a lottery that result in c^* for E_0 and c_* for the complementary event $E - E_0$. It is postulated that for any event E_0 a number $\bar{\alpha}$ exists, such that the two lotteries are equivalent. This number $\bar{\alpha}$ is denoted by $P(E_0)$. In other words, this model can only be used if we feel that the real life problem can be compared with a hypothetical lottery. Before giving the formal postulates we have to introduce the following notation.

Let E be a finite set of real world events. Denote by $X \times Y$ the set of outcomes of a two-dimensional hypothetical experiment; the outcomes are restricted to the generalized interval $[0,1]$. On the Cartesian product $E \times X \times Y$ all sorts of lotteries can be defined. Such lotteries are denoted by

$$\ell(e, x, y) = c^* \quad \text{if} \quad e \subset E_0,\ x \subset X_0 \text{ and } y \subset Y_0$$
$$= c_* \quad \text{otherwise}$$

where $E_0 \subset E$ and where X_0 and Y_0 are sub-intervals of $[0,1]$. Apparently, if $E_0 = E$, the lottery is defined on the hypothetical canonical experiment only; if $X_0 \times Y_0 = X \times Y$, the lottery is defined on the set of real life events.

As far as preferences exist in the class of lotteries, it is supposed that these preferences are transitive. This implies that a partial ordering is defined on the class of lotteries. If ℓ_2 is not strictly preferred to ℓ_1, this is denoted by $\ell_2 \precsim \ell_1$, and we have

P.PRS.1. If $\ell_1 \precsim \ell_2$ and $\ell_2 \precsim \ell_3$, then $\ell_1 \precsim \ell_3$.

If both $\ell_1 \precsim \ell_2$ and $\ell_2 \precsim \ell_1$, this equivalence will be denoted by $\ell_1 \sim \ell_2$. Let us now return to the lotteries that are defined on the canonical experiment. Evidently, it is sensible to choose a canonical experiment that is a simple as possible. It is supposed that the canonical experiment is characterized by a two-dimensional rectangular probability distribution.

This is reflected by

P.PRS.2a. If $\ell_1(e, x, y) = \begin{cases} c^* & \text{if } e \in E_0 \text{ and } (x, y) \in X_1 \times Y_1 \\ c_* & \text{otherwise} \end{cases}$

$$\ell_2(e, x, y) = \begin{cases} c^* & \text{if } e \in E_0 \text{ and } (x, y) \in X_2 \times Y_2 \\ c_* & \text{otherwise} \end{cases}$$

if moreover the area of $X_1 \times Y_1$ is equal to the area of $X_2 \times Y_2$, then $\ell_1 \sim \ell_2$.

P.PRS.2b. If $E_0 = E$ and the area of $X_1 \times Y_1$ is strictly larger than $X_2 \times Y_2$ then $\ell_1 \succ \ell_2$.

The restriction $E_0 = E$ in this postulate has to be made because the decision-maker could attach the value 0 to the chance $E_0 \times X_0 \times Y_0$ whatever area $X_0 \times Y_0$ is chosen.

The third postulate asserts that the lottery defined on $E_0 \times X \times Y$ may be compared with a lottery defined on $E \times X \times Y_0$. In other words lotteries on real life events may be compared with lotteries on the hypothetical canonical experiment.

P.PRS.3. For any subset $E_0 \in E$ there exist a value $P[E_0]$ such that

$$\ell_1(e, x, y) = \begin{cases} c^* & \text{if } e \in E_0 \text{ and } (x, y) \in X \times Y \\ c_* & \text{otherwise} \end{cases}$$

and

$$\ell_2(e, x, y) = \begin{cases} c^* & \text{if } y \in [0, P[E_0]] \text{ and } (e, x) \in E \times X \\ c_* & \text{otherwise} \end{cases}$$

are equivalent.

The fourth and last postulate is needed to assure that the property of additivity is not only valid for canonical chances but also for the probability function $P[E_0]$.

P.PRS.4. Let Q be any one of E, X, Y; let q be the corresponding one of e, x, y; let $\{Q_1, \ldots, Q_n\}$ be a partition of Q; for i = 1, ..., n, let ℓ_i' and ℓ_i'' be lotteries upon $E \times X \times Y$ independent of the value of q; let

$$\left.\begin{array}{l} \ell'(e, x, y) = \ell_i'(e, x, y) \\ \ell''(e, x, y) = \ell_i''(e, x, y) \end{array}\right\} \quad \text{if } q \in Q_i$$

Then, if $\ell_i' \sim \ell_i''$ for all i, $\ell' \sim \ell''$ holds.

From these postulates one can derive the existence of a unique finitely additive subjective probability measure on the class of events $\{E_0\}$.

The importance of this approach lies in the fact that it makes explicit a line of thought that is implicitly used by Savage. Numerical subjective probabilities only exists if comparisons with hypothetical lotteries can be made. It is quite interesting to mention that the idea of a canonical experiment has been used by Hemelrijk [7] in an entirely different context. The latter defines the concept "aselector" which is an <u>objective</u> canonical experiment. This concept is used as a basis to define an objective probability concept in a frequency context.

With the canonical experiment we end up not very far from our starting point: real life decision problems are replaced by betting situations (section 4) or reference lotteries. Is there a reason to prefer the latter above the former? Conceptually, there is a difference: Bets are considered for given events, and the consequences (the odds) can be chosen from a range of possible consequences. Reference lotteries are concerned with only two consequences c^* and c_*, and the event is chosen from the range [0,1]. When a whole range of consequences is involved, then we cannot duck the question about the utility function of those consequences. The <u>assumption</u> that utilities are linear for small amounts of money makes it clear that, in its essence, subjective probabilities are defined simultaneously with utilities. Things are different for reference lotteries: It is only assumed that $c^* > c_*$, and the relative value of c^* and c_* is of no importance (It is remarkable that Pratt, Raiffa and Schlaifer have defined subjective probabilities and utilities simultaneously, although this is not necessary in their approach). This theoretical advantage must be paid for with a price in practical applications: To operate with reference lotteries is more difficult than to imagine bets.

7. Bayesianism in the Mixed Approach and in De Finetti's Representation Theorem.

Up till now, we have considered the meaning of subjective probabilities in terms of consistent betting behaviour and in terms of consistent preference orderings for acts. The question has been left open why a person puts his odds in some specific way, or why a person has a specific preference ordering for his acts. The consistency requirements alone are far from sufficient for the derivation of a unique probability measure. The problem is of great practical importance, because the task of assessing probabilities is usually a very difficult one; we would like to have some guiding principles at our disposal to help us in our task. Such guiding principles have not necessarily the same status as the consistency requirements: Assuming that a person is willing to play along, inconsistencies can be punished by proposing combinations of bets. On the other hand, bets that are in clear conflict with common sense could - by chance - be rewarded, instead of being punished. If, for instance, a coin has been flipped a large number of times, and if about half of the times the outcome heads has been observed, then it is common sense that a bet on the next outcome should be on even odds. Nevertheless, someone might be willing to bet with odds 1:9, and this might turn out to be profitable.

Betting odds reflect our ideas about the phenomenon in question. If these ideas are very complex, we cannot expect guiding principles to exist that reduce the problem of the assessment of probabilities to a simple algorithm. If, for instance, the future market share for a newly-developed product is to be estimated, degrees of belief will reflect our knowledge about market shares for analogous products, our experience of the market in other countries, our ideas about the activities of competitors, and so on. The marketing manager will weigh all such information in his mind, without making use of algorithms. The construction of guiding principles serves a more modest purpose; it is meant for those cases where the available information is fairly homogeneous. For instance, one might think of a procedure to estimate future sales on the basis of the sales figures for the last ten years.

In its most elementary form, the problem is the following: If in n trials k successes have been observed, how can this kind of information be used for the assessment of the probability that a success will occur in the next trial? In textbooks on subjective probability, the following example is often used: Let there be an urn with red and white balls. The experiment consists in removing a ball, observing its colour, putting it back into the urn and shuffling the balls around, after which the experiment can be repeated. Let the proportion of red balls in the urn be equal to p. Now let us assume that we can take a quick glance at the contents of the urn before the experiment starts. Although we do not know the value of p, we have some idea of its possible value. We may construct a (subjective) probability distribution for p with density f(p). This is called the prior distribution. Now, if it is assumed that the experiment is a Bernouilli process, then for any given p the probability of k successes in n trials can be written as $\binom{n}{k}p^k(1-p)^{n-k}$. Weighting this result with the degrees of belief for the different values of p, the (subjective) probability of k successes in n trials can be written as

$$P[k/n] = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} f(p)dp \qquad (1)$$

In this way, a multidimensional (subjective) probability distribution is defined upon p and sequences of outcomes. But then, the ordinary rules for conditional probabilities can be applied. It is easily seen from Bayes' theorem that

$$f(p|\tfrac{k}{n}) = \frac{P[\frac{k}{n}|p]f(p)}{P[\frac{k}{n}]} = \frac{p^k(1-p)^{n-k}f(p)}{\int_0^1 p^k(1-p)^{n-k}f(p)dp} \qquad (2)$$

where k/n stands for an observed relative frequency of k red balls in n trials. These observations have changed our prior ideas about the value of p. The new distribution, the density of which is given in (2), is called the posterior distribution. In the same way, bets on relative frequencies will be revised in the light of already observed relative frequencies k/n:

$$P[\tfrac{t}{s}|\tfrac{k}{n}] = \frac{P[\frac{t}{s} \cap \frac{k}{n}]}{P[\frac{k}{n}]} = \frac{\binom{n}{k}\binom{s}{t}\int_0^1 p^{k+t}(1-p)^{s+n-(k+t)}f(p)dp}{\binom{n}{k}\int_0^1 p^k(1-p)^{n-k}f(p)dp} =$$

$$= \int_0^1 P[\tfrac{t}{s}|p]f(p|\tfrac{k}{n})dp \qquad (3)$$

where the last equality is easily verified by means of (2). This way of incorporating knowledge from observations has some interesting features, that are in agreement with common-sense requirements. First, it is shown that an extra observation of red increases the probability that red will be observed in a new trial. In formula, we want to prove that

$$P[R|\tfrac{k+1}{n+1}] \geq P[R|\tfrac{k}{n}] \qquad (4)$$

This inequality can be rewritten as

$$\frac{\int_0^1 p^{k+2}(1-p)^{n-k}f(p)dp}{\int_0^1 p^{k+1}(1-p)^{n-k}f(p)dp} \geq \frac{\int_0^1 p^{k+1}(1-p)^{n-k}f(p)dp}{\int_0^1 p^{k}(1-p)^{n-k}f(p)dp} \qquad (5)$$

and the proof is a direct application of Schwarz's inequality.

The second, and more important, property has to do with the limiting behaviour of the posterior distribution if the observed relative frequency tends to a fixed limit. Indeed, if we are concerned with a Bernouilli process, we may expect that the observed relative frequency will tend to the "true" value of the parameter. Suppose that k/n tends to a fixed number r for $n\to\infty$, then it is easily seen that the mass of (2) is concentrated around r, provided that f(p) does not vanish around r. Therefore, if the possibility that p = r is not excluded beforehand by our choice of the prior distribution f(p), we know that in the long run the posterior distribution will disclose the "true" value of p. This, of course, affects our actual betting behaviour. If we bet on the outcome red in single trials, (3) can be written as

$$P[R|\tfrac{k}{n}] = \int_0^1 pf(p|\tfrac{k}{n})dp \qquad (6)$$

and because $f(p|k/n)$ tends to a one-point distribution for $k/n \to r$, we find

$$P[R|\tfrac{k}{n}] \to r \text{ for } \tfrac{k}{n} \to r, \text{ provided that } f(p) \neq o \text{ around } p = r$$

In other words, in the long run we shall use betting odds that correspond with the "true" value of p.

It is noteworthy that the learning model, as described above, is developed on the basis of two assumptions. First, it is assumed that we are able to construct a prior distribution f(p) by having the opportunity to glance over the contents of the urn. This immediately leads to a number of objections:

- The frequentist will say that the subjectivity makes the learning model inappropriate for scientific application. According to him, science must exclude things like "personal ideas" or "subjective feelings". The subjectivist will answer that, in the long run, objective truth will suppress the subjective prior distribution, as follows from (6). If a large number of observations are available, the subjectivist is as close to the truth as the frequentist, whereas the subjectivist can also apply the learning model to situations where only a limited number of observations are available. Therefore, the learning model described here is more general than the frequentist's inference procedures.

- People who are interested in practical applications will remark that it is only very seldom that we have the opportunity to glance over the contents of the urn. Very often, we know next to nothing of the "true" value of p. To overcome this problem, attempts have been made to construct "vague priors". Such attempts have not yet resulted in generally accepted procedures, and perhaps this may be impossible.

- As soon as we try to generalize the learning model, we run into problems of interpretation. Suppose that the model is used for a coin-tossing experiment, then p can no longer be interpreted as a proportion of a finite population, but it must be considered as a probability in the frequentistic sense. As we shall show at the end of the next section, this leads to considerable complications.

The second assumption that has been made in the construction of the learning model is the one about Bernouilli trials. The model, as used so far, combines subjective probabilities (the prior distribution) and frequentistic probabilities (Bernouilli trials for given p). This duality is reflected in the term "mixed approach" with which we denote the learning process described here. The mixture of two different kinds of probabilities is at the least confusing, and should be clarified. One way to solve the problem is to stick to subjective probabilities as the essence of the learning process, where these sujective probabilities are assessed as if we have to do with Bernouilli trials. This, however, takes away the basis of rationality from the learning process. Why should we act as if we have to do with Bernouilli trials, when we are not willing to admit that anything like frequentistic probabilities exists? Another way of solving the problem would be the explicit recognition of two different kinds of probabilities. This has been advocated by Carnap. Unfortunately, he has not elaborated this concept. A third way out is offered by the Finetti, who reduces the prior distribution to a mathematical artefact that can be derived from betting behaviour in special circumstances. Suppose that the subjective probabilities for an outcome R after n observations are influenced by the relative frequencies of successes, but that the order in which successes and failures have been observed does not influence the subjective probabilities, then it can be proved that there is some function f*(p) such that

$$P[R|\tfrac{k}{n}] = \frac{\int_0^1 p^{k+1}(1-p)^{n-k} f^*(p)dp}{\int_0^1 p^{k}(1-p)^{n-k} f^*(p)dp} \qquad (7)$$

for all k and n. This is the essence of de Finetti's representation theorem. The existence of the "prior distribution" f*(p) is not prior to the betting behaviour, but it is a consequence of the betting behaviour if we deal with information in some specific way. The proof of the theorem is rather complicated, and the reader is referred to [4].
Formally, the distribution f*(p) has all the properties of a prior distribution, but its interpretation is quite different.

The situation where the ordering of successes and failures is considered to be irrelevant is of great practical importance. Such a situation is covered by the concept of exchangeability. More precisely, if a sequence of events, "successes" and "failures", is such that the probability of k successes and m failures in k + m experiments is dependent upon k and m only, this sequence is called exchangeable. In the next section we shall see how Carnap takes exchangeability as a starting-point for his approach (together with the ordinary rules for the calculation of probabilities), and to what extent he succeeds in the development of further guiding principles fot the choice of a "rational betting system".

8. The Concept of Normative Probabilities

Up till now we have met two different interpretations of the concept of probability: the frequency approach as, for instance, supported by Von Mises and the subjectivistic approach as supported by De Finetti, Savage and others. It is important to realize that the notion of probability is used in different ways in these two cases. Carnap [2] was well aware of this difference and made it clear that the word probability is not restricted to a single meaning. At least two different conceptions can be distinguished, in Carnap's terminology probability_1 and probability_2.

We shall now give a description of these conceptions; if we use the notion of probability in the sense of probability_2, we have in mind the empirical fact of stabilizing relative frequencies. In other words, frequentists use the notion of probability in the sense of probability_2. On the other hand, if we use the notion of probability in the sense of probability_1, we have the following conception in mind:

(1) the probability that the next toss of a coin will be heads given a certain amount of information.

(2) the probability of an event as reflected in a fair betting quotient.

(3) probability_1 can sometimes be interpreted as the <u>estimation of a relative frequency</u>. Carnap [2] clarifies this with the aid of the following example. Imagine an urn filled with 60 red and 40 white balls. One ball is drawn from the urn. Assume that two players bet on the outcome of this draw. What is their fair betting quotient? According to Carnap it is 6 : 4. In other words, the fair betting quotient is determined by the proportion of white (red) balls in the urn. Carnap defends this solution with the following argument: if we use this rule for 100 simultaneous bets (one bet on each ball) the final balance is foreseeable and equal to zero. Now it goes without saying that, if we do not know the proportion of white balls in the urn, the fair betting quotient will be equal to the estimate of it.

From this point on, Carnap takes two steps that are essential for the rest of his work on probabilities. Firstly, he states that the two concepts of probability are not necessarily in conflict. One should see the two different notions as complementary. The use of a frequentistic model for any real world situation forces us to develop rules of inference, which connect the unknown values of the parameters in the model with the observed data, and also connect statements about future (finite) sets of observables with the frequentistic model if the latter is completely known. These rules of inference do not have the character of probability_2, although they might be based upon probability_2 statements. This can be illustrated by the following example: If probabilities_1 are measured such that events that are considered impossible have a probability_1 value 0, then a frequentist's rule of inference like "events with a very low probability are supposed to be impossible" can be translated as "events with very low probability_2 values have probability_1 value 0". Of course "very low" should be specified in a statement like this. In section 2 we have seen already that assessment of probability_1 values in this way rules out the possibility of a calculus of probability_1 statements. This is in conflict with the requirement that probability_1 statements should reflect "rational" betting behaviour, when combinations of bets are permitted.

A better way to proceed is the following: If the probability_2 value of an event equals p, then the probability_1 value that is given to the occurrence of this event in a single trial is also equal to p. So, if it is known that a coin has the physical property that it falls heads half the times that it is tossed in an (infinite) sequence of trials, then it is rational to bet with even odds on the next outcome.

It is evident that a rule like the above is outside the scope of subjectivists. According to them, the only rationality rules for betting behaviour are the ones that require coherence (excluding betting odds that result in a certain loss or a certain gain for finite combinations of bets). A broad-minded subjectivist who acknowledges the existence of probability_2 might say that most people will bet in accordance with this rule, but he will never admit that people should bet in accordance with it. It is at this point that Carnap takes the second step that is of fundamental importance for his work. He tries to develop rules of rational betting behaviour, and from his earlier work it appears that he is striving for a unique way of deriving probability_1-values. Because subjective feelings have nothing to do with such a system, Carnap uses the term "degree of confirmation" rather than "degree of belief". Hence, $c(H|E)$ denotes the degree of confirmation that hypothesis H will be true in the light of evidence E, and this can be considered as a normative assessment of the degree of belief that H will turn out to be true, given the evidence E. In his earlier work, Carnap tried to build a system of degrees of confirmation by means of the logical structure of the language in which H and E are described. Hence, the term "logical probabilities" was used for these kinds of probabilities. It appeared, however, that the logical structure of the language in which H and E are described is not a sufficient basis for the derivation of unique degrees of confirmation, and therefore we will use the term normative probabilities, rather than logical probabilities, for all systems that impose upon degrees of belief rationality requirements which exceed the coherence requirements laid down by the subjectivists.

The rule that $probability_1$ values should be chosen such that they are equal to $probability_2$ values might appeal to our intuition; there is, however, a serious objection to such a rule. $Probability_2$ values can never be observed, because they are defined for infinite sequences of observations. We would need a rule of inference to derive $probability_2$ values from the evidence E. But then, it seems better to base degrees of confirmation directly upon the observable evidence E, instead of making a detour in terms of $probability_2$. In other words, Carnap tries to develop a rational way of using evidence E for betting on hypothesis H. We may call this a rational learning process.

Henceforth, we shall base our considerations mainly upon Stegmüller [20], who has given an elaborate survey of Carnap's later (partly still unpublished) ideas. The reader is also referred to Carnap [3] where special attention is given to Carnap's concept of a "model". Before turning to the axioms, we must first define the kind of hypotheses H and evidence E for which the axioms are assumed to hold true. Let $\{a_i\}$ be a (finite or countable) set of individuals. For a coin-tossing experiment this is the set of trials. Each individual will turn out to be characterized by a finite set of attributes $\{A_1, A_2, \ldots, A_n\}$ where each member A_i belongs to a family F_i of mutually exclusive attributes. It is assumed that each family of attributes F_i contains at least 2 and at most a countable number of different attributes. If a statistical experiment is performed in which people's height and weight are measured, there are two families of attributes. The number of attributes within each family generally depends on the refinement of the measuring method. If height is measured to the nearest centimeter, and if we feel certain that everybody's height is between 1 and 400 centimeters, there are 400 different attributes in this family. In the coin-tossing example there is only one family of attributes, containing the two attributes "heads" and "tails". Now a <u>model</u> is defined as a complete specification of all individuals in terms of n attributes (each of which belongs to the corresponding family of attributes). Note that a model is <u>not</u> defined as the <u>true</u> specification, but rather as a <u>possible</u> specification of all individuals. For the coin-tossing example, for an infinite number of trials, <u>any</u> infinite sequence of the outcomes "heads" and "tails" can be considered as a model. Next, consider the set of all models. Subsets can then be defined in which for any specified individual a_i the j^{th} attribute for the m^{th} family is fixed. The proposition "individual a_i has the j^{th} attribute from the m^{th} family" is called an atomic proposition, and is formally identified with the subset of models for which this is true. In the coin-tossing example, the atomic proposition "the third trial results in "heads" " is formally identified with the set of all sequences of outcomes "heads" and "tails" that have "heads" in the third place. In this way, set-theoretical rules can be applied to propositions. Now let E_{at} represent the set of all atomic propositions. The set E_{mol} of all molecular propositions is defined as the algebra, generated by E_{at} on the set of all models, and the set E of all propositions is defined as the σ-algebra, generated by E_{at} on the set of all models. In the coin-tossing example, propositions like "the first toss will result in "heads" and the third toss will result in "tails" " and "in the first n tosses the outcome "heads" will occur k times or less" are molecular. Propositions like "all trials with even number will result in "heads" " belong to E, but are not molecular. In this example, any model (that is, any completely specified infinite sequence of outcomes heads and tails) belongs to E, but does not belong to E_{mol}.

It is noteworthy that, in the example given above, the set of all models in which the relative frequency of heads tends to some fixed limit $r \leq p$ (where p is given) does not belong to E. In other words, the hypothesis that we are dealing with some $probability_2$ value less than p is not a proposition in the sense of the definition of E. Such a hypothesis is typical for frequentistic estimation procedures, and an extension of the set of admissible propositions so that this kind of hypothesis could be included might prove to be fruitful.

It is assumed that in $c(H|E)$ both the hypothesis H and the evidence E are propositions in the sense defined above. This does not imply that $c(H|E)$ is defined for all propositions. One might be interested in only one family of attributes, leaving aside a second family of attributes. In that case, the set E would be too broad, and we should restrict ourselves to some set $E' \subset E$. Moreover, it is assumed that the evidence E is not logically impossible.

The axioms that should be fulfilled by $c(H|E)$ are given here in two parts. We shall start with the axioms that are also met within the work of subjectivists (either as axioms or as assumptions). Then we shall consider Carnap's λ-family of confirmation functions. Thereafter, other axioms that reflect rationality will be considered.

Probability$_1$ should reflect rational betting behaviour, and it is natural to require the property of coherence just as in the case of subjective betting behaviour. As we have pointed out already in section 4, the requirement of coherence is a requirement of rationality that is not necessarily imposed upon a person in actual betting behaviour (it could well be that, for instance, a complementary bet does not exist in reality). In our search for rationality requirements, the axioms that are derived from the requirement of coherence come in rather naturally. Therefore, we state:

P.C.1. $c(H|E) \geq 0$
P.C.2. $c(E|E) = 1$
P.C.3. $c(H|E) + c(-H|E) = 1$, where $-H$ is the complement of H
P.C.4. If $E \cap H \neq \emptyset$, then $c(H \cap H'|E) = c(H'|E \cap H).c(H|E)$.

These are the fundamental axioms for any probability theory (not including σ-additivity, however).

P.C.5. For any two molecular propositions H and E such that $H \cap E \neq \emptyset$, the requirement $c(H|E) > 0$ must hold.

This postulate ensures strict coherence, or in words: Combinations of bets that result in either zero profit or in a loss are excluded. At first sight, this requirement seems to be rational indeed. Nevertheless, it is a bit tricky. Suppose that somebody sincerely believes that when a coin is tossed twice the only possibilities are twice heads or twice tails. Both these possibilities are considered equally likely. He will then accept a bet at even odds that twice heads will occur, and he will also accept a bet at even odds that twice tails will occur. The combination of these two bets will lead to zero gain if twice heads or twice tails is actually observed. He will lose both his stakes if once heads and once tails is actually observed. The person will not think of himself as being irrational in his betting behaviour, because he does not believe that the loss will actually occur! To most people, the example of coin tossings supports the necessity of the requirement of strict coherence, because they are well aware of the possibility that once heads, once tails might obtain. A slight modification in the example could change this attitude completely. Suppose that we throw two goldfish into an unknown chemical liquid. Then it seems to be quite rational to assume that either both fish will die or both fish will survive. The idea that some fish will die while other fish survive may seem rather irrational. Only if a phenomenon is recognized as being stochastic, is the requirement of strict coherence plausible! We shall return to this subject when we are dealing with Carnap's λ-family of confirmation functions.

P.C.6. If H and E are given; if the individuals i and j are interchanged in both H and E, such that H' and E' are derived, then $c(H|E) = c(H'|E')$.

According to this postulate, our degree of confirmation that heads will occur in the second trial, given the evidence that heads occurred in the first trial, is equal to the degree of confirmation that heads have occurred in the first trial, given the evidence that heads occurred in the second one. The way in which the individuals have been ordered does not play a relevant role in the determination of degrees of confirmation. This postulate is equivalent to the Finetti's concept of interchangeability. Again, there is more in this postulate than one would think at first sight. If our evidence consists of 1000 observations, all even numbers being heads and all uneven numbers being tails, the degree of confirmation that heads will occur at the 1001st trial is equal to the degree of confirmation that heads will occur at the 1002nd trial. Now a rational person would perhaps give higher probability to the latter, because the trial has again an even number, but this way of reasoning is excluded by P.C.6. In the case that we feel certain that one outcome is physically independent of any other outcome, we would indeed treat the 1001^{st} trial in the same way as the 1002^{nd} trial; the remarkable fact that the outcomes heads and tails alternate would be put down to chance. The postulate requires that we remain convinced that the outcomes are <u>physically</u> independent, whatever (finite) evidence we may have to counter this conviction.

It might be useful to pay some attention to the connection between physical independence and statistical independence. Frequentists who think that an experiment consists of physically independent trials will put

$$P_f(\text{heads}/1000 \text{ times heads in } 1000 \text{ trials}) = P_f(\text{heads}) = p.$$

The multiplication rule holds, because for example

$$P_f(\text{twice heads in 2 trials}) = P_f(H_2/H_1).P_f(H_1) = p^2.$$

The subjectivists, and also Carnap, give a different interpretation of the probability concept. P_s reflects our degree of belief (confirmation), and our belief will be influenced by experience. Therefore, in general,

$$P_s(\text{heads}/1000 \text{ times heads in } 1000 \text{ trials}) \neq P_s(\text{heads})$$

and

$$P_s(\text{twice heads in 2 trials}) = P_s(H_2/H_1).P_s(H_1) \neq \{P_s(H)\}^2.$$

Physical independence of trials is reflected in the assumption of exchangeability (or, equivalently, in P.C.6), and is quite different from statistical independence. On the other hand, statistical independence can also be met within the subjectivistic approach. If our degrees of belief are such that $P_s(H|E) = P_s(H)$, where H and E stand for hypothesis and evidence respectively, then there is formally statistical independence. We shall give two examples, each of which has its own pecularities. First, consider the coin-tossing example. Suppose that no trials have been made so far, and that a bet on even odds is considered fair. Then $P_s(H) = \frac{1}{2}$. Now if 1000 tosses have been made, and exactly 500 of these have resulted in heads, then it seems rather natural to bet with even odds again on the next outcome being heads. Although there have been quite a large number of observations, these have not influenced our degree of belief. It should be noted, however, that the situation after 1000 tosses differs drastically from the original situation. Although the degrees of belief are the same, the sensitivity to new evidence will presumably have decreased substantially. If the new evidence consists of ten times heads in a row, then we may expect that $P_s(H_{11}/10 \text{ times H in ten tosses})$ will be much closer to 1 than $P_s(H_{1011}/510 \text{ times H in 1010 tosses})$.

The second example of statistical independence in the subjectivistic theory is of more importance. We can make an a priori choice that a certain kind of evidence does not influence our degrees of belief for a certain kind of hypothesis. Think of the following experiment: A coin is taken from a bag, it is tossed, and after the observation it is put back in the bag. Now suppose that there are two kinds of coins in the bag, silver and nickel ones. The observation consists of the result of the toss and the metal of the coin (in the terminology introduced above, there are two families of attributes). If we firmly believe that the outcome of the toss teaches us nothing about the metal, the two families of attributes can be called (statistically) independent: $P_s(H|E) = P_s(H)$ whenever H is a hypothesis about the metal only that (one or more) coins are made of and E is evidence about the outcome of the tosses only. The multiplication rule then, for instance, takes the form

$$P_s(\text{SH}/3 \text{ times SH and 4 times ST and 7 times NH and 6 times NT}) =$$
$$P_s(\text{S}/7 \text{ times S and 13 times N}).P_s(\text{H}/10 \text{ times H and 10 times T})$$

where S and N stand for silver and nickel, and H and T for heads and tails. It may be questioned whether such a way of learning from experience is wise. Suppose that in a large number of trials most nickel coins fall heads, whereas most silver coins fall tails. Then it seems unwise to neglect the outcome of the toss of the next coin in the determination of our degree of belief that this coin will be silver. The use of this kind of evidence has been blocked however by the a priori choice of (statistical) independence for the two families of attributes.

The axioms, as given so far, are also found in the subjectivistic approach when exchangeability exists. We are still a long way off from a rational betting system. Other restrictions must be imposed for that purpose. Two lines of attack can be followed: We can analyse the general structure, or we can try to find rules of rationality for relatively easy problems, like throwing a die. We shall start with the latter. Suppose there is only one family of attributes, the number of attributes being k. Our experience may consist of s observations ($s \geq 0$). Let s_i denote the number of times that attribute i has been observed ($0 \leq s_i \leq s$). Let $P_i a_j$ denote that individual j has attribute i; if there is no danger of confusion, we shall suppress a_j, so P_i must be read as "some given individual has attribute i" or (sometimes) "the next individual to be observed has attribute i". Now we are interested in the construction of degrees of confirmation $c(P_i|S)$ where S describes the observations already made. Bearing in mind the example of throwing a die, it seems, for some cases at least, rational to require:

R.1. The degree of confirmation for hypothesis P_i is only dependent on the relative frequency s_i/s which has been observed (the way in which the other $s-s_i$ observations are distributed over the other attributes is of no importance).

First, let us consider the situation that no observations have yet been made ($s = 0$). We define $c(P_i|Z) = \gamma_i$, where Z stands for certainty, in other words, Z will say that the only experience we have is irrelevant. Sometimes we shall require:

R.2. All γ_i-values are equal.

Because $\sum_{i=1}^{k} c(P_i|Z) = 1$ it follows immediately from R.2. that $\gamma_i = 1/k$. This requirement may seem rational for the die-throwing example if the die is of regular form; one might also use this requirement if one does not know anything about the mechanism of the experiment, falling back on the principle of insufficient reasoning. Henceforward, however, we shall drop R.2.

Requirement R.1. makes it possible to use a notation that is very economical. $c(P_i|k/n)$ should be read as the degree of confirmation that an individual which has not yet been observed has attribute i, given that in n observations attribute i has been observed k times. Now consider the situation where one observation has been made. From axioms P.C.1. through P.C.6. and from requirement R.1. we can derive that

$$\frac{c(P_i|0/1)}{\gamma_i} = \eta \text{ is constant for all } i \tag{1}$$

This can be seen as follows: According to R.1. it makes no difference for $c(P_i|0/1)$ what attribute actually has been observed, as long as it is not attribute i. So choose any attribute $j \neq i$ then

$$\frac{c(P_i|0/1)}{\gamma_i} = \frac{c(P_ia_2|P_ja_1)}{\gamma_i} = \frac{c(P_ia_2 \cap P_ja_1|Z)}{\gamma_i c(P_ja_1|Z)} = \frac{c(P_ja_1|P_ia_2) \; . \; (P_ia_2|Z)}{\gamma_i \; . \; (P_ja_1|Z)} =$$

$$= \frac{c(P_ja_1|P_ia_2).\gamma_i}{\gamma_i.\gamma_j} = \frac{c(P_j|0/1)}{\gamma_j}$$

The first equality sign is needed to come to a more explicit notation, the second and third equality signs are immediate consequences of P.C.4., the fourth equality sign is derived from the definition of γ_i and γ_j. Using the definition of η, we find

$$c(P_i|0/1) = \eta\gamma_i \quad \text{for } i = 1, \ldots, k \tag{2}$$

To derive $c(P_i|1/1)$ we recognize the fact that the new individual will certainly have one of the k attributes, so the degrees of confirmation for the different P_j's must add up to one. If for attribute i the observed relative frequency is 1/1, then for any attribute $j \neq i$ the observed relative frequency must be 0/1, so

$$c(P_i|1/1) + \sum_{j \neq i} c(P_j|0/1) = 1 \tag{3}$$

Using the result (2), this may be rewritten as

$$c(P_i|1/1) = 1 - \sum_{j \neq i} \eta\gamma_j = 1 - \eta \sum_{j \neq i} \gamma_j = 1 - \eta(1 - \gamma_i) \tag{4}$$

The last equality sign holds because $\sum_{i=1}^{k} \gamma_i = 1$.

With (2) and (4), all degrees of confirmation are determined for the case of one observation. The results (2) and (4) can be combined if we define

$$\lambda = \frac{\eta}{1-\eta} \,. \tag{5}$$

Simple arithmetic shows that

$$c(P_i | s_i/s) = \frac{s_i + \lambda\gamma_i}{s + \lambda} \tag{6}$$

for s = 0 or s = 1, and $0 \leq s_i \leq s$. The next step is the extension of (6) for all values $s \geq 0$ and $0 \leq s_i \leq s$. It can be proved that, <u>if the family of attributes contains at least three attributes</u>, if P.C.1. through P.C.6. hold, and if R.1. must be satisfied, then (6) must hold for all $s \geq 0$. The proof is given by means of complete induction. Unlike Stegmüller [20], we shall construct the proof directly on basis of (6). Suppose (6) has been proved for the case that there are s observations, then we have to prove (6) if the number of observations is increased by one. Let s_i denote the frequency with which attribute i has been observed in s observations, so $\sum_{i=1}^{k} s_i = s$.

We shall use S as a shorthand notation for the k-tuple of observed relative frequencies $\{s_1, \ldots, s_k\}$. First, we state that the individual a_{s+1} must have one of the k attributes with certainty, so

$$c(P_i a_{s+2} | S) = \sum_{j=1}^{k} c(P_i a_{s+2} \cap P_j a_{s+1} | S) =$$

$$= \sum_{j=1}^{k} c(P_i a_{s+2} | S \cap P_j a_{s+1}).c(P_j a_{s+1} | S) =$$

$$= \sum_{j \neq i} c(P_i a_{s+2} | S \cap P_j a_{s+1}).c(P_j a_{s+1} | S) +$$

$$+ c(P_i a_{s+2} | S \cap P_i a_{s+1}).c(P_i a_{s+1} | S).$$

Because (6) has been assumed to be true for s observations, it can be used for those terms where the number of observations does not exceed s. Using (6), and rewriting the other terms in an easier notation, we find

$$\frac{s_i + \lambda\gamma_i}{s + \lambda} = \sum_{j \neq i} c(P_i | \frac{s_i}{s+1}).\frac{s_j + \lambda\gamma_j}{s + \lambda} + c(P_i | \frac{s_i + 1}{s + 1}).\frac{s_i + \lambda\gamma_i}{s + \lambda}$$

Because $c(P_i|\frac{s_i}{s+1})$ is not dependent on j, it can be taken before the summation sign. Because

$$\sum_{j\neq i} \frac{s_j + \lambda\gamma_j}{s+\lambda} = \frac{s - s_i + \lambda(1-\gamma_i)}{s+\lambda}$$

we find, after simplification

$$1 = \frac{s - s_i + \lambda(1-\gamma_i)}{s_i + \lambda\gamma_i} \cdot c(P_i|\frac{s_i}{s+1}) + c(P_i|\frac{s_i+1}{s+1}) \tag{7}$$

For the next step, suppose that the $(s+1)^{th}$ observation turns out to have attribute i. Because we know with certainty that the $(s+2)^{th}$ individual must have one of the k attributes, we may say

$$c(P_i|\frac{s_i+1}{s+1}) + \sum_{j\neq i} c(P_j|\frac{s_j}{s+1}) = 1$$

and this may be rewritten as

$$c(P_i|\frac{s_i+1}{s+1}) + \sum_{j=1}^{k} c(P_j|\frac{s_j}{s+1}) - c(P_i|\frac{s_i}{s+1}) = 1 \tag{8}$$

Using (8) to eliminate the term $c(P_i|\frac{s_i+1}{s+1})$ in (7), we find after some easy arithmetic

$$c(P_i|\frac{s_i}{s+1}) = \frac{s_i + \lambda\gamma_i}{s+\lambda} \cdot \sum_{j=1}^{k} c(P_j|\frac{s_j}{s+1}) = K \cdot \frac{s_i + \lambda\gamma_i}{s+\lambda} \tag{9}$$

where K is the shorthand notation for the sum that is involved. Using (9) in equation (8), we find

$$c(P_i|\frac{s_i+1}{s+1}) + K - K \cdot \frac{s_i + \lambda\gamma_i}{s+\lambda} = 1 \tag{10}$$

From (9) and (10), all degrees of confirmation for the $(s+2)^{th}$ individual are derived, up to an unknown factor K. To determine K, we recognize the fact that, if $s \geq 1$, at least one attribute will have occurred with a relative frequency $< \frac{s}{s}$ (let this be attribute i), whereas at least one attribute will have occurred with relative frequency $\frac{0}{s}$ (let this be attribute 1). Now, if we suppose that we started with another set S' of s observations, such that $s_i' = s_i + 1$, $s_1' = s_1 - 1$ and $s_j' = s_j$ for all other attributes, we would find, according to (9)

$$c(P_i|\frac{s_i'}{s+1}) = \frac{s_i' + \lambda\gamma_i}{s+\lambda} \cdot \sum_{j=1}^{k} c(P_j|\frac{s_j'}{s+1}) = K' \cdot \frac{s_i' + \lambda\gamma_i}{s+\lambda}$$

so

$$c(P_i|\frac{s_i + 1}{s+1}) = K' \cdot \frac{s_i + 1 + \lambda\gamma_i}{s+\lambda}. \quad (11)$$

Now, if the number of attributes in the family is at least equal to 3, we are able to prove that K' = K. Consider for that purpose equation (9). This equation must hold whatever values are given to s_j ($j \neq i$), as long as s_i remains constant. Therefore, K must be insensitive to all such changes. For an attribute j, (9) would have the form

$$c(P_j|\frac{s_j}{s+1}) = K \cdot \frac{s_j + \lambda\gamma_j}{s+\lambda}$$

and using the same line of thought, K must be insensitive to all changes as long as s_j remains constant. But then, K is insensitive to any change, as long as $\sum_{j=1}^{k} s_j = s$. K' has been derived by taking $s_i' = s_i + 1$ and $s_1' = s_1 - 1$, so K' = K.

The above way of reasoning cannot be followed if only two attributes are contained in the family of attributes. If $s_1 + s_2 = s$, and if $s_1' = s_1 + 1$, then we find automatically $s_2' = s_2 - 1$. There is no value s_i that remains constant, as was required in the proof above.

The rest of the proof consists of simple arithmetic. From (10) and (11), putting K' = K, we easily find

$$K = \frac{s+\lambda}{s+\lambda+1}$$

and using this in (9) and (10) we find

$$c(p_i|\frac{s_i}{s+1}) = \frac{s_i + \lambda\gamma_i}{s + 1 + \lambda}$$

$$c(P_i|\frac{s_i + 1}{s+1}) = \frac{s_i + 1 + \lambda\gamma_i}{s + 1 + \lambda}$$

With this result, the proof is completed.

In the case where only two attributes are contained in the family of attributes, (6) cannot be derived from R.1. and the axioms P.C.1. through P.C.6. The requirement R.1. plays an important part in the proof, and if there are only two attributes, R.1. is no longer a restriction: It is automatically fulfilled. It is, however, plausible to require that (6) should be valid in this case as well.

Consider again the experiment, consisting of throwing a die. If we only pay attention to an even or uneven outcome, such an experiment could be considered as an analogue to the coin-tossing experiment. For the die we have found from R.1.

$$c(\text{uneven}|s_1 \cap s_2 \cap s_3 \cap s_4 \cap s_5 \cap s_6) = c(P_1|\frac{s_1}{s}) + c(P_3|\frac{s_3}{s}) + c(P_5|\frac{s_5}{s}) =$$

$$= \frac{s_1 + \lambda\gamma_1}{s + \lambda} + \frac{s_3 + \lambda\gamma_3}{s + \lambda} + \frac{s_5 + \lambda\gamma_5}{s + \lambda} = \frac{s_{\text{uneven}} + \lambda\gamma_{\text{uneven}}}{s + \lambda}$$

This formula has the same specification as (6), and it contains only the attributes "uneven" and its negation ("even"). There are other, more direct reasons for choosing the confirmation function (6) in the case of only two attributes, but that part of the discussion will be postponed. First, we shall consider the properties of (6) more closely; in illustrating several properties, we shall freely assume that (6) holds also for the special case where only two attributes are involved.

First, we notice that $c(P_i|\frac{s_i}{s}) \geq 0$ for all $s \geq 0$ and $0 \leq s_i \leq s$, and from this it appears directly that $\lambda \geq 0$ in (6). Next, (6) can be rewritten as

$$c(P_i|\frac{s_i}{s}) = \frac{s_i + \lambda\gamma_i}{s + \lambda} = (\frac{s}{s + \lambda})(\frac{s_i}{s}) + (1 - \frac{s}{s + \lambda})\gamma_i$$ and because

$\lambda \geq 0$ we find that the degree of confirmation $c(P_i|\frac{s_i}{s})$ is a convex combination of the a priori value γ_i where no experience is available yet, and the observed relative frequency $\frac{s_i}{s}$. This property can be interpreted in the following way: If no experience is available yet, we start by giving P_i a degree of confirmation γ_i. After observing the relative frequency s_i/s, we shall change our opinion in the direction of the observed relative frequency; the rapidity with which we change our opinion is characterized by the parameter λ. Consider the two extreme values of λ. If $\lambda = \infty$, then $c(P_i|s_i/s)$ degenerates to γ_i for <u>any</u> s_i/s. In other words, if $\lambda = \infty$ our prior ideas are so strong that they do not change whatever evidence we get. In a coin-tossing experiment, we might be absolutely convinced that the coin is fair. Even if we find heads 100 times in 100 tosses, we might still think that this event is due to chance, and continue to believe that the coin is fair. It would require more stubbornness than one can imagine, if the idea that the coin is fair is not changed <u>whatever</u> evidence is brought forward. The choice $\lambda = \infty$ does not seem rational, and should be excluded.

The other extreme occurs if we choose $\lambda = 0$. The confirmation function (6) then becomes $c(P_i|s_i/s) = s_i/s$ for $s \geq 1$ and $c(P_i|0/0) = \gamma_i$. As soon as any evidence is acquired, this evidence determines our degree of confirmation completely. At first sight, this might seem to be a rational way of proceeding if our starting position were one of absolute ignorance. The value γ_i could be found from the principle of insufficient reasoning (leaving aside the problems that are connected with this principle), and when evidence has been acquired the value of γ_i is of no consequence any more. Nevertheless, the choice $\lambda = 0$ is excluded by Carnap, because he feels it to be irrational that $c(P_i|1/1) = 1$ and $c(P_i|0/1) = 0$ as would follow from $\lambda = 0$. Only one positive observation would make us feel certain that the next outcome would again be positive, whereas one negative observation would lead us to the belief that it is impossible that the next outcome turns out to be positive. Technically, the choice $\lambda = 0$ is excluded by the axiom P.C.5. From this axiom, it appears that $c(P_i|s_i/s)$ is <u>strictly</u> positive for all $s \geq 0$ and $0 \leq s_i \leq s$, so

$$c(P_i|s_i/s) = \frac{s_i + \lambda\gamma_i}{s + \lambda} > 0$$

and putting $s_i = 0$, $s \geq 1$ and $\gamma_i > 0$ we find $\lambda > 0$. The question whether or not we should admit the choice $\lambda = 0$ is essentially the same as the question whether or not P.C.5. is justified.

In the case that there are only two attributes, the axioms P.C.1. through P.C.6. satisfy the requirements of De Finetti's representation theorem (and add very little to it). Therefore, we know that there must be some prior distribution f(p) such that $c(P_i|\frac{s_i}{s})$ can be written as

$$c(P_i|\frac{s_i}{s}) = \frac{\int_0^1 p^{s_i+1}(1-p)^{s-s_i} f(p)dp}{\int_0^1 p^{s_i}(1-p)^{s-s_i} f(p)dp} .$$

As a matter of fact, <u>any</u> prior distribution f(p) that is positive on the interval [0,1] is consistent with the axioms P.C.1. through P.C.6. If we want to use formula (6), we are no longer free to choose f(p). It is easily verified that (6) is formally equivalent to a mixed approach in which the prior distribution f(p) is beta-distributed with $\lambda\gamma_i$ and $\lambda(1-\gamma_i)$ as degrees of freedom: From

$$f(p) = \frac{\Gamma(\lambda)}{\Gamma(\lambda\gamma_i)\Gamma(\lambda - \lambda\gamma_i)} \cdot p^{\lambda\gamma_i - 1}(1 - p)^{\lambda(1 - \gamma_i) - 1}$$

We easily find

$$c(P_i|\frac{s_i}{s}) = \frac{\int_0^1 p^{s_i+1}(1-p)^{s-s_i} f(p)dp}{\int_0^1 p^{s_i}(1-p)^{s-s_i} f(p)dp} =$$

$$= \frac{\int_0^1 p^{s_i+\lambda\gamma_i}(1-p)^{s-s_i+\lambda(1-\gamma_i)-1} dp}{\int_0^1 p^{s_i+\lambda\gamma_i-1}(1-p)^{s-s_i+\lambda(1-\gamma_i)-1} f(p)dp}$$

$$= \frac{\Gamma(s_i+\lambda\gamma_i+1)\ \Gamma(s-s_i+\lambda(1-\gamma_i))/\Gamma(s+\lambda+1)}{(s_i+\lambda\gamma_i)\ \Gamma(s-s_i+\lambda(1-\gamma_i))/\Gamma(s+\lambda)} = \frac{s_i+\lambda\gamma_i}{s+\lambda} .$$

and this is exactly the form of (6).

Now let us return to the extreme values of λ. If $\lambda \to \infty$, the prior distribution tends to a one-point distribution with all its mass on the point $p = \gamma_i$. If $\lambda \to 0$, the prior distribution tends to a two-point distribution, with its mass concentrated on the points $p = 0$ and $p = 1$. We have seen already that the choice $\lambda = 0$ corresponds to a situation where the a priori degree of confirmation γ_i is completely dominated by the observations already made. This seems to be typical for the non-informative situation. Is it possible to interpret the corresponding prior distribution in terms of lack of information?

Raiffa and Schlaifer answer this question in the negative; they argue as follows (see [18], page 65):
" (...) we find that the beta distribution does not approach a proper limiting distribution: namely a two-point distribution with a mass of γ_i on $p = 1$ and a mass $(1 - \gamma_i)$ on $p = 0$[1]. Now this limiting distribution cannot in <u>any</u> sense be considered "vague". On the contrary, it is completely <u>prejudicial</u> 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 annihilate the mass at $p = 0$, and a single failure will annihilate 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 continued 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 <u>not</u> strictly 0. Long experience with a particular production process or

----

1) The notation is adapted to the context of this chapter. The underlinings are part of the quotations.

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".

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 a stochastic phenomenon. The situation cannot be called non-informative. A real non-informative situation is one for which no analogous situation exists. 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 the points $p = 0$ and $p = 1$. It is only after the observation of both successes and failures, that we consider the phenomenon as a stochastic one. Formally, if the relative frequency s_i/s has been observed with $s_i < s$ the posterior distribution $f_i(p|\frac{s_i}{s})$ is beta-distributed with $s_i + \lambda\gamma_i$ and $s - s_i + \lambda(1 - \gamma_i)$ degrees of freedom, and for λ tending to 0 this posterior distribution has its mass spread over the whole interval $[0,1]$.

The argument put forward by Raiffa and Schlaifer, that one would have to be very convinced before putting all probability mass on the points $p = 0$ and $p = 1$, is misleading. Of course, if an enormous amount of information were available, this might be a sufficient reason for using this prior distribution. On the other hand, it is not necessary for the use of this prior distribution that such information is available. We may choose a learning process that makes us as open-minded as possible for observations in the future, in other words we leave the prior degree of confirmation γ_i as soon as possible and choose $\lambda = 0$.

Let us now return again to the properties of the confirmation function (6).

We have seen already that, when (6) holds, the degree of confirmation $c(P_i|s_i/s)$ must be between the prior degree of confirmation γ_i and the observed relative frequency s_i/s. The observed relative frequency draws the degree of confirmation away from the original value γ_i, but the degree of confirmation never exceeds s_i/s. This property seems to be rational, and it is not guaranteed by the axioms P.C.1. through P.C.6. As a counter example, take a prior distribution with density

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

Then

$$\gamma_i = \int_0^1 p(3 - 12p(1 - p))dp = \tfrac{1}{2}.$$

However,

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

and, clearly, this value is not in the interval [1/3,1/2]. On the other hand, it can be shown that there are prior distributions that fulfil the above requirement, although they are not beta-distributed. The requirement that $c(P_i|s_i/s)$ must be between γ_i and s_i/s is not sufficient for the derivation of (6). The properties of (6) are much stronger than this, but they are in the same line of thought. Suppose that the relative frequency s_i/s has been observed, then the posterior value $c(P_i|s_i/s)$ can be considered as a new starting point, so $c(P_i|s_i/s) = \tilde{\gamma}_i$. Now suppose that t new observations have been made, with t_i "successes". Then we may require that the degree of confirmation is drawn away from $\tilde{\gamma}_i$, but does not exceed t_i/t. In formula:

R.3. For all s, $t \geq 0$, $0 \leq s_i \leq s$ and $0 \leq t_i \leq t$, it holds that

$$c(P_i|\frac{t_i + s_i}{t + s}) \text{ is between } \frac{t_i}{t} \text{ and } c(P_i|\frac{s_i}{s}).$$

If (6) holds, requirement R.3. is necessarily fulfilled, as can easily be verified from

$$c(P_i|\frac{t_i+s_i}{t+s}) = \frac{t_i+s_i+\lambda\gamma_i}{t+s+\lambda} = (\frac{t}{t+s+\lambda}).(\frac{t_i}{t}) + (1 - \frac{t}{t+s+\lambda}).\, \frac{s_i+\lambda\gamma}{s+\lambda}$$

Is R.3. in addition to P.C.1. through P.C.6. sufficient to derive (6) from it in the case of two attributes? For specific choices of γ and η this can be proved to be true. I do not know whether the proof can be given in general. It is, however, hardly worth spending much attention on it. We may choose (6) because it is an elegant form that satisfies R.3., whereas most other proceedings (that is, choices of a prior distribution $f_i(p)$) would violate this requirement.

Now that we have analysed the proporties of (6), we turn to the question of the cases in which it is not rational to use (6). For a die-throwing experiment, requirement R.1. seems to be rational. If, however, we are interested in the age that a child starts to read, the possible outcomes could also be 1 through 6, but the observed relative frequency of children who begin to read at the age of 5 will affect our opinion that the next child to be observed will begin to read at age 6 more than the relative frequency of children who begin to read at age 4. Requirement R.1. is not fulfilled, and we need some kind of "distance function" to describe in what way we want to learn from experience. The efforts at defining a distance function have not so far been very fruitful, and it may seriously be questioned whether such a distance function can be defined on other than personalistic grounds.

There are three other axioms mentioned in [20], that must be quoted here. First we need an axiom that excludes the use of $\lambda = \infty$ in (6). This is done by

P.C.7. $\lim_{s\to\infty} \{c(P_i|s_i/s) - s_i/s\} = 0$

This axiom implies that, in the long run, the degree of confirmation is almost completely determined by the observed relative frequency s_i/s, although it is not necessary that s_i/s tends to a limit itself! In this way the case $\lambda = \infty$ is excluded in (6), because $c(P_i|s_i/s)$ would be constant and independent of the observed relative frequency s_i/s.

To make the theory fit with traditional statistics, we might require

P.C.8. $c(H|Z)$ is a σ-additive function with E as domain of definition for H.

In this, Z stands for certainty. The question whether or not we should require σ-additivity is still open to discuss, and we shall not go into this discussion here.

Finally, we need an axiom that connects the learning system for, say, a thousand throwings of a die with a learning system for a million throwings of a die; that connects a learning system for a die with the 6 possible outcomes distinguished with a learning system where only the outcomes even and uneven are distinguished; that connects a learning system for k families of attributes with a learning system for k-m families of attributes. It is required that

P.C.9. (a) If H and E are expressed in a context of k trials, and if H' and E' have the same meaning in a context of k + n trials, then $c(H|E) = c^*(H'|E')$.
(b) If H and E are expressed in one family of attributes, and if H' and E' are derived from H and E if a number of attributes are taken together, then $c(H|E) = c^*(H'|E')$.
(c) If H and E are expressed in a context of r families of attributes, and if H' and E' are derived from H and E by neglecting certain families, then $c(H|E) = c^*(H'|E')$.

The first part of this axiom presents no difficulties. The second part is intuitively clear, but it excludes the introduction of the principle of insufficient reasoning: Suppose that we are interested in the experiment of throwing a die, then the principle of insufficient reasoning would lead us to the assumption that the degrees of confirmation for the 6 different outcomes should be equal; on the other hand, if we are interested in the outcomes "6" and "not-6" respectively, the principle of insufficient reasoning would lead us to degrees of confirmation $\frac{1}{2}$ and $\frac{1}{2}$, instead of 1/6 and 5/6. It seems reasonable, however, that the principle of insufficient reasoning is sacrificed for those cases where we are able to define a larger family of attributes properly. The third part of the axiom is the most troublesome. Suppose that we are interested in two families of attributes, the first being whether or not a driver caused an accident last year, the second being the sex of the driver. Now, if we can neglect the sex of the driver without changing the confirmation values, then apparently we are not able to learn from the observed correlation between sex and accidents. P.C.9. (c) prescribes that one family of attributes can never teach us anything about another family of attributes, that there is thus complete independance, whatever our experience may be. This is in conflict with ordinary statistical practice.

Now that we have described Carnap's approach to the probability concept, we can compare the different ways in which the probability concept is worked out. The frequentists try to deduce a system of statistical inference by means of probability$_2$-statements. In our opinion, such an attempt is rather fruitless, unless explicit attention is paid to the fact that statistical inference is in fact nothing but the use of degrees of belief in a very specific way (the very improbable is conceived as impossible). More interesting is the comparison between the subjectivistic approach and Carnap's approach. The subjectivists claim that there is no verifiable reason for imposing other restrictions on betting behaviour than the consistency rules. Carnap is trying to find criteria for rational betting behaviour. Both approaches make use of the Bayesian principle. Now, if there is (subjective) information, it seems rather stupid to forego this information because it does not fit into Carnap's system, so the subjectivists are right in following their ideas in this specific situation. On the other hand, if no (subjective) information is available, then the only thing that we can do is to establish a learning system, and for that purpose special requirements can be made, as for instance requirement R.3. If a person uses a prior distribution $f_i(p)$, it will be difficult to interpret the meaning of it. Is it the reflection of intuitive ideas, or is it a reflection of the way in which he wants to react to new information? To answer this question, we can best look at the possible interpretation of p. If p can be given a clear interpretation, as in the case where an urn contains red and white balls, and the person has had the opportunity of glancing over the contents of the urns, then a prior distribution on p has a well-defined, though subjective meaning. If, on the other hand, p reflects the unknown probability that some irregularly-formed coin will fall heads up, then a prior distribution $f_i(p)$ can only describe the way in which we want to learn from experience. The intermediate case is the most difficult one: We have vague knowledge of the true parameter p, and we want to learn from experience as it is acquired. In this case, there seems to be only one solution: The prior distribution, vague as it is, should be changed in such a way that the requirements for a learning process are not violated; on the other hand, the learning process should be chosen in such a way, that the prior ideas are not essentially violated.

The conclusion to this chapter is the following:
According to the personalistic approach there exists complete freedom in betting situations to put one's odds at will, as long as the requirement of coherence is fulfilled. If this approach is applicated to situations where information of repetitive character is present, other restrictions (like R.3.) can be introduced to arrive at the learning process that is in accordance with intuition. Bayesian statisticians should not stop at the personalistic approach, but, on the contrary, they should start from this point on, defining what kind of restrictions can reasonably be superimposed under what kind of circumstances.

9. Some Concluding Remarks.

In retrospect, are there essential differences in the assumptions that are made by frequentists and subjectivists? We may not compare the frequentistic theory with Bayesian information processing. It is rather the frequentistic inference procedure that must be used for a comparison. If we do so, it becomes quite clear that:

- subjectivity plays a role in the frequentistic inference rule, as well as in the subjectivistic approach. The frequentistic inference rule ("the very improbable is consdiered impossible") can be translated as "if the frequentistic probability of an event is known to be very low, then the degree of belief that such an event will have taken place in a specific situation is equal to 0". Of course, degrees of belief that are defined in this way do not fulfil the rules of calculation for probabilities. As we have stated already in section 3, there is no need for such rules of calculation if we are dealing with mass phenomena.
- subjectivists using prior distributions for Bayesian information processing learn from observed relative frequenties. But then, it stands to reason that they implicitly believe in stabilizing relative frequencies, just as the frequentists. What other reason could there be for them to look at such data?

The essential difference is in the domain of application. The frequentists restrict themselves strictly to mass phenomena, and the subjectivists - as far as they use prior distributions - have a much wider domain that they cover. The preocedures are different because they cover different grounds. What are the conclusions from the foregoing for the economist who is facing some problem where uncertainty is involved? The answer is by no means unique: The economist may have to deal with a problem that can properly be described as a mass phenomenon. In that case he will no doubt use the frequentistic theory. At another time he may be confronted with, say, the estimation of the sales for the next few years for a completely new product. Subjective probabilities can then be of help. Rather often he will be required to make estimations on basis of a small number of observations. It is quite clear that he cannot use a pure frequentistic theory. He might be inclined to use an adapted form of the frequentistic theory (for instance, using a 60% confidence level). This, however, seems a rather dangerous way of solving his problem. The inference rule for frequentists is in its essence to define away very small probabilities, and it is rather dangerous to use techniques from this area for probabilities that are no longer small. On the other hand, there is a serious drawback as to the use of subjective probabilities. If there is not much prior information then there is an element of arbitrariness in the assessment of prior distributions. For readers of the economist's report it will be difficult to find out what influence this arbitrariness has upon the results. It is of course true that the reader can find out by calculating things over with his own prior distribution, but this will be a rather timeconsuming work. Moreover, this work is often not worthwhile, because the results are anyhow rather uncertain. But then, it would simplify the communication if unique standard techniques can be used. The complete freedom of the pure subjectivist leads to undesireable individualism, and we must look for normative probabilities as an intermediate between pure frequentism and pure subjectivism. The problem then becomes how to find such normative probabilities. Two aspects are relevant. Arithmetical simplicity and conceptual menaingfulness. As to the latter, one might think of the kind of criteria that we have met in section 8. If conceptually meaningful criteria do no lead to a unique normative probability measure (as is the case in Carnap's confirmation function where λ is not specified), then it may be worth to complement them with artificial criteria. For Carnap's confirmation function, this could be a standard choice for the value of λ, for instance $\lambda = 0$, the advantage of standardization might well outweigh the disadvantage in the lack of interpretation.

<u>List of references</u>

(1) Bernouilli, Jacob: Ars conjectandi, Basel (1713)
(German translation: Wahrscheinlichkeitsrechnung, Leipzig (1899)).

(2) Carnap, R.: Logical foundations of probability, 2nd ed., Chicago (1962).

(3) Carnap, R.: A basic system of inductive logic, part I, in: Studies in Inductive Logic and Probability, Vol. I, ed. by R. Carnap and R.C. Jeffrey, Berkeley - Los Angeles - London (1971).

(4) Feller, W.: An introduction to probability theory and its applications, Vol. II, New York - London - Sydney (1966).

(5) De Finetti, B.: La Prévision: ses lois logiques, ses sources subjectives, Annales de l'Institut Henri Poincaré, 7 (1937).

(6) De Finetti, B.: Probability, induction and statistics, London - New York - Sidney - Toronto (1972).

(7) Hemelrijk, J.: Back to the Laplace definition, Statistica Neerlandia, 22 (1968).

(8) Halmos, P.R.: Measure theory, 2nd printing, New York (1950).

(9) Jeffreys, H.: Theory of probability, 3rd edition, Oxford (1961).

(10) Kolmogoroff, A.N.: Grundbegriffe der Wahrscheinlichkeitsrechnung, Berlin (1933).

(11) Koopman, B.O.: The axioms and algebra of intuitive probability, Annuals of Mathematics, Ser. 2, 41 (1940).

(12) Koopman, B.O.: Intuitive probabilities and sequences, Ser. 2, 42 (1941) Annuals of Mathematics.

(13) Kyburg, H.E.: Probability and the logic of rational belief, Middletown (1961).

(14) De Laplace, P.S.: Théorie analytique des probabilités, Paris (1812).

(15) Von Mises, R.: Probability, Statistics and Truth, London (1957).

(16) Von Mises, R.: Mathematical theory of probability and statistics, New York - London (1964).

(17) Pratt, J.W.; H. Raiffa and R. Schlaifer: The foundations of decision under uncertainty: an elementary exposition, Journal of the American Statistical Association 59 (1964).

(18) Raiffa, H. and R. Schlaifer: Applied statistical decision theory, Boston (1961).

(19) Savage, L.J.: The foundations of statistics, New York (1954).

(20) Stegmüller, W.: Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band IV, Personelle und Statistische Wahrscheinlichkeit, Heidelberg - New York (1973).

REPORTS 1977

7700 List of Reprints, nos. 179-194; List of Reports, 1976.

7701/M "Triangular - Square - Pentagonal Numbers", by R.J. Stroeker.

7702/ES "The Exact MSE-Effiency of the General Ridge Estimator Relative to OLS", by R. Teekens and P.M.C. de Boer.

7703/ES "A Note on the Estimation of the Parameters of a Multiplicative Allocation Model", by R. Teekens and R. Jansen.

7704/ES "On the Notion of Probability: A Survey", by E. de Leede and J. Koerts.