Satosi Watanabe

Information-Theoretical Aspects of Inductive and Deductive Inference

Abstract: By a straightforward application of Bayes' theorem of probability, the behavior is discussed of the credibilities (inductive probabilities) of competing hypotheses as functions of an increasing body of relevant empirical data. It is shown how the effect of a priori credibilities persists in the evaluation of credibilities in general, except in the important limiting cases investigated. An "inverse H-theorem" is mathematically demonstrated, according to which the entropy function defined in terms of the credibilities shows a net decrease in time. This decrease is not necessarily monotonous in an individual case, but is monotonous in the "expected" behavior of the inductive entropy function. Three machine-simulation experiments of inductive inference on the IBM 704 are described. The first two concern the classical problem of guessing the ratio of white and black balls in an urn. The third experiment concerns guessing a hidden pattern obeyed by a sequence of binary numbers.

1. Introduction and methodological remarks

• *A. Need for machine simulation of inductive process*

Efforts in simulating human cognitive processes on computing machines have recently shown surprising progress, as exemplified by Gelernter's geometry-problem-solving machine[1] and Samuel's checker-playing program.[2] One of the most interesting features found in some of these advanced programs is that the computing machine which was primarily designed to carry out deductive inference seems to be executing some task which constitutes part of inductive inference. This may easily be observed, for instance, in the case of Samuel's checkers program,[2] in which various criteria (hypotheses) for winning are evaluated on the basis of empirical data obtained by the machine during games. Admittedly, the available criteria are first chosen in this case by men and given to the machine.

These newer developments in programming suggest that a practical need will be felt more and more acutely in the future for a well-founded mathematical method of executing as much as possible of what is called inductive inference, including hypotheses testing. A quest of such a method may provide also some clue to the basic methodological questions, such as whether a machine can perform the entire inductive process, including the so-called "creative" work which is often considered a necessary ingredient of inductive inference. Without pretending to give any ultimate answers to such basic questions, the present paper is intended to be a contribution to mathematical analysis of inductive inference as a whole, making necessarily some reference also to deductive inference. Those readers who are not particularly interested in nonmathematical, methodological arguments are invited to pass from this point to the beginning of Section 2.

• *B. Conditions for a mathematical model of induction*

The present paper may be considered as a possible mathematical explication of the procedures involved in inductive inference. No claim is made that human intelligence actually follows or should follow precisely the prescriptions derivable from this explication. Neither is it contended that this form of explication covers all the subtle aspects of inductive inference. It is only a mathematical model. The reader will find, however, that many of the important features of inductive inference are adequately represented in this model theory. In the following, we shall mention ten important features of inductive inference which any adequate theory of inductive inference should incorporate in some way or another, and which the present mathematical model indeed does. Admittedly, these ten conditions may not be sufficient but are certainly necessary.

(1) *Role of deductive inference.* Inductive inference contains, as a necessary ingredient, a constant comparison of the deductive consequence from a hypothesis with the experiment. Accordingly, the model theory of inductive inference must permit deductive inference to play a corresponding role within its framework.

(2) *Logical refutation by counterexample.* This is the most elementary step in inductive inference, in which a hypothesis is disqualified when the hypothesis excludes the occurrence of a certain event (observed datum) while actual experience shows that this forbidden event in fact occurs (is observed). This is the only part of the inductive process where deductive logic is the sole arbiter. It is surprising that many philosophical papers are still written at this elementary level of consideration. Nevertheless, this process of logical refutation must be included in any theory of inductive inference.

(3) *Continuous measure of preferential confidence on hypotheses.* The essential theoretical difficulty concerning the process of inductive inference stems mainly from the fact that there usually exist a great number of, indeed often infinitely many, hypotheses which are not logically refuted by the available evidence, and which are not necessarily unanimous regarding the outcome of an observation which is not yet made known. For this reason, inductive inference is often declared to be "logically" ill-founded. It should be noted, however, that in actual human inductive inference, we usually place preference on one hypothesis to another even though both are not logically refuted. There seem to be two sources for such preference. One of these two sources will be discussed under (5) below. The other source is the fact that the body of evidence is capable not only of refuting or not refuting a hypothesis but also of furnishing a continuous degree of support to a hypothesis. For instance, suppose that there are two hypotheses H_1 and H_2, and that the first hypothesis H_1 allows the occurrence of two events D_1 and D_2 with nonvanishing probabilities, while the second hypothesis H_2 allows only the occurrence of D_2. Suppose further that the actual body of evidence consists only of D_2. Then neither H_1 nor H_2 is logically refuted, but we have to place preference on hypothesis H_2, for it better fits the experimental data. This degree of preference will depend on the probabilities placed on D_1 and D_2 by H_1, and also the preference of H_2 must be stronger if the number of D_2 in the body of evidence is larger. This simple example is sufficient to show that we must be allowed to attach a continuous measure of preferential confidence, or *credibility*, or inductive probability,[3] to each of the competing hypotheses. We shall agree on the convention that the value of credibility equal to unity (the largest possible value) would mean that the hypothesis is a "law" and the value equal to zero would mean that the hypothesis is totally incompatible with experience. A logically refuted hypothesis will have credibility zero, but there may be cases where the credibility of a hypothesis tends to zero in the limit with an increasing body of evidence, even though the hypothesis is not logically refuted.

(4) *Successive approach.* The essence of scientific method resides not in discovering an absolute truth but in successive improvement of knowledge. This is true whether the term "improvement" means the applicability of a theory to a broader domain of experience, or the capability of a theory in yielding more precise agreement with the experimental measurement within a given domain of experience, or a better fit of the predicted frequency distribution of various results with the experimental frequency. This basic nature of scientific method must be reflected in any theory of inductive inference. Thus it seems natural to require that the theory be based on a procedure by which we "modify" or "improve" the evaluation of credibility in the measure as the body of evidence accumulates. In our model, we use Bayes' Theorem as the basis for the formula for improvement of the evaluation of credibilities. It is not contended that this is the only justifiable way to establish such a formula, but this approach certainly has many attractive features.

(5) *Effect of judgment from broader experience.* A test of hypotheses must be defined by some observational operation, and such a test must be instrumental in the above-mentioned successive improvement of the evaluation of the credibilities. However, in this actual evaluation, enough flexibility must be left to accommodate the consideration originating from a broader field of experience, of which the test in question represents only a small part. Such a flexibility is needed to permit a unifying structure of a "theory" covering a wide area of experience. Such broader consideration also serves greatly to invent new hypotheses as well as to degrade useless hypotheses before the test.

(6) *Absolute certainty of validity of hypothesis denied.* No hypothesis should be declared to be a law (i.e., credibility unity) on the basis of a finite number of observed data. This is closely connected with the fact that it is impossible to derive a conclusion (or a hypothesis) for an infinite number of cases from the experience of a finite number of cases.

(7) *Existence of law with objective validity.* Notwithstanding the remark (6) above, we cannot deny the existence of a law (probabilistic or deterministic) governing a limited area of experience, for such denial would amount to renouncing scientific quest in general. Corresponding to this situation, it must be guaranteed that some hypothesis, whether or not already considered, reaches credibility unity in the limit where the size of the body of evidence becomes infinitely large. And this selection of hypothesis must be independent of any preconceived judgment, except in the case where there is more than one "equivalent" hypothesis.

(8) *Distinction between credibility and confirmability.* As stated under (3), credibility is the degree of preferential confidence. In other words, it is a relative weight among the competitive hypotheses. As was seen in (5), the credibility is bound to be influenced by the experience at large, except in the limiting case discussed under (7). Distinct from credibility, there must be a certain measure of the degree to which a test (which is a series of the same type of observation) confirms a hypothesis individually taken, completely independent of the other hypotheses and of the experience outside the test in question. This degree of confirmation will be called *confirmability* and will be normalized so that it becomes unity when the confirmation becomes "perfect." Although credibility and confirmability are conceptually distinct, a high confirmability must tend to increase the credibility.

(9) *Room for new hypotheses.* Usually we cannot from the outset think of all the possible hypotheses to cope with a certain series of experiments. On the contrary, a new hypothesis usually occurs to a scientist after he has accumulated a certain amount of experimental facts. Therefore, the model theory of induction must be such that we can add a new hypothesis at any stage of the process of induction and let it compete with the other hypotheses which have already been considered. In this case, of course, past experience must also be reviewed in the light of the new hypotheses.

(10) *Anti-ergodicity and inverse H-theorem.* Inductive inference is a process such that the distribution of weights (credibilities) becomes increasingly concentrated on a decreasing number of cases (hypotheses) no matter how widely one distributes the weights initially. Loosely speaking, this is contrary to the tendency of an ergodic stochastic chain in which, no matter on what case one might put the weight initially, the distribution of weights gradually spreads out to all the cases, which are "connected" to the initially chosen case. Correspondingly, there must be a theorem showing the tendency opposite to the H-theorem if the (inductive) entropy is defined suitably with the aid of credibilities. As the H-theorem shows an increase (or nondecrease) of the entropy with time, in a certain sense of the average value, the inverse H-theorem can be expected to show a decrease (or non-increase) of the inductive entropy with the growth of experience, in a certain sense of the average. It seems that the inverse H-theorem has deep philosophical implications, as has the usual H-theorem (entropy principle of thermodynamics) proved to have in the past. But we refrain from philosophizing on this matter here.

• C. *Method based on Bayes' theorem*

Now a few more remarks explaining how these requirements are actually satisfied by the method based on Bayes' theorem. To make these explanations briefly, let us first define some convenient symbols. Let $\mathfrak{D}$ be the set of all possible outcomes of a well-defined observation of a well-defined kind of phenomena; let $\mathfrak{B}^{(\nu)}$ be a sequence of the outcome of the past ν observations of the kind defined in connection with $\mathfrak{D}$; and let $\mathfrak{H}$ be the set of hypotheses under consideration that allow one to calculate the deductive probability of obtaining any one of the elements of the set $\mathfrak{D}$. Thus $\mathfrak{B}^{(\nu)}$ is the body of evidence so far obtained. If two hypotheses give the same deductive probability distribution with respect to a certain subset of $\mathfrak{D}$, we shall say that these two hypotheses are probabilistically equivalent with respect to this subset. We denote by $q^{(\nu)}(H)$ the credibility we attach to hypothesis H (a member of $\mathfrak{H}$) on the basis of $\mathfrak{B}^{(\nu)}$. We might have included the cases where $\mathfrak{H}$ contains an infinite number of hypotheses, but we confine the discussion to the finite case, since the mathematical complications due to infinity may mar the more important issues.

The use of the continuous measure $q^{(\nu)}(H)$ satisfies Requirement (3) above. Bayes' theorem is, in accordance with Requirement (4), a mathematical formula permitting calculation of $q^{(\nu)}(H)$ from $q^{(\nu-1)}(H)$ with the help of the ν^{th} observed datum and of the deductive probabilities attributed to the ν^{th} datum by various H's. The use of deductive probabilities here corresponds to Requirement (1). Bayes' theorem is such that if a hypothesis H gives zero deductive probability to the ν^{th} observed datum, then $q^{(\nu)}(H)$ of this hypothesis becomes zero no matter what value $q^{(\nu-1)}(H)$ may have. This shows that the process of logical refutation, Requirement (2), is automatically performed by Bayes' theorem. However, Bayes' theorem, being a prescription for obtaining $q^{(\nu)}(H)$ from $q^{(\nu-1)}(H)$, cannot determine, as far as ν is finite, the actual values of the credibilities, leaving arbitrary a set of constants, $q^{(0)}(H)$, which may be interpreted as the *a priori* credibilities of hypotheses. Here, the term *a priori* is to be understood as "not solely determined by $\mathfrak{B}$", i.e., the empirical data of the observation as specified in the definition of $\mathfrak{D}$. We can capitalize on this fact to accommodate Requirement (5) by letting $q^{(0)}(H)$ represent the judgment from a broader experience. The only exception to this general rule (that $q^{(0)}(H)$ are needed for evaluation of the values of $q^{(\nu)}(H)$ with a finite ν) is the case where the process of logical refutation leaves one and only one hypothesis. This case, however, happens very seldom. The indispensability of the a priori credibilities becomes particularly manifest when there are two hypotheses which are probabilistically equivalent with respect to the outcome included in $\mathfrak{B}^{(\nu)}$. In this case, the ratio of the credibilities of these two hypotheses becomes simply the ratio of the a priori credibilities attached to these hypotheses, which may be evaluated by other tests or by a broader consideration.

An immediate consequence of the above observation is that no hypothesis can be crowned as the law on the basis of a finite size of the empirical data, agreeing with Requirement (6). Indeed, any credibility evaluated on the basis of a finite $\mathfrak{B}^{(\nu)}$ depends on the a priori credibilities, and a change in the a priori credibility causes a change in *a posteriori* credibility at a finite stage ν. The evaluation of these a priori credibilities may be considered as a product of another inductive process of broader range or higher level, but this inductive process, again being based on a finite experience, can never give definite values of these a priori credibilities.

On the other hand, the present paper also allows one to draw some comforting conclusions, in compliance with Requirement (7). If a set of probabilistically equivalent hypotheses is considered as a single hypothesis with regard to $\mathfrak{D}$, then we can show that, as the empirical data accumulate, the credibility of one of the hypotheses approaches unity, i.e., $q^{(\nu)}(H)\to 1$ with $\nu\to\infty$, and the credibilities of the remaining hypotheses approach zero, no matter how the a priori credibilities are given. (For a more rigorous discussion of the condition for $q^{(\nu)}(H)\to 1$, see Section 4.) It can also be shown that the credibility of each hypothesis becomes less sensitive to the a priori credibility as the size of the empirical data increases. It should be noted that two equivalent hypotheses with regard to a specific observation defining $\mathfrak{D}$ may not be equivalent for another specific observation. This will help differentiate two hypotheses and determine the preference between them. It may also be noted that even if it is guaranteed that $q^{(\nu)}(H)\to 1$ with

$\nu \to \infty$ for a particular H, the value of $q^{(\nu)}(H)$ for a finite ν depends on the a priori credibilities of all hypotheses in $\mathfrak{H}$; hence an appropriate evaluation of a priori credibilities is always a great help when ν is finite.

Section 3 assumes the existence of the limit: $q^{(\infty)}(H)$ for $\nu \to \infty$, under a very lenient condition about $\mathfrak{B}^{(\nu)}$, and concludes that the limit must be 0 or 1, if a family of probabilistically equivalent hypotheses is counted as one hypothesis. On the other hand, in Section 4, it is assumed that the frequency in $\mathfrak{B}^{(\infty)}$ of various events belonging to $\mathfrak{D}$ has a definite distribution, and it is then shown that this limit $q^{(\infty)}(H)$ indeed exists and is 0 or 1, provided that a hypothesis whose confirmability (see below) becomes unity at the limit is included in $\mathfrak{H}$.

As regards Requirement (8), it should be noted that a hypothesis always gives a deductive probability for the occurrence of each possible event included in $\mathfrak{D}$, and that this probability distribution can be compared with the actual frequency of occurrence of various events in $\mathfrak{B}^{(\nu)}$. The present paper gives in Section 5 a nice measure for the degree of agreement of these two probability distributions (one predicted by H and the other empirical), thus offering a method to determine the confirmability required in (8). This confirmability is, of course, different from the credibility, but it will be shown that if the confirmability of a hypothesis becomes unity (its maximum value) in $\mathfrak{B}^{(\infty)}$, then this hypothesis is bound to be granted $q^{(\infty)}(H) = 1$. But on the other hand, if none of the hypotheses reaches confirmability unity, then a hypothesis whose confirmability is the highest will reach credibility unity. This may be called a "law" since it is the "best" available hypothesis. This is the reason why we use $q^{(\infty)}(H) = 1$ as the definition of a law. A stricter definition of a law would be to require both credibility and confirmability to be unity. It may be noted that too strong reliance on the confirmability at a finite ν is misleading, since, in contrast to the credibility, the confirmability can accidentally become unity at a finite ν even though its value at $\nu \to \infty$ is not unity.

Objections have been raised to the application of Bayes' theorem to the problem of inductive inference. An answer to some of these objections will be presented in a concrete example in Section 4C. The gist of the view proposed in this paper in this connection is the extreme flexibility which is allowed in evaluating the a priori credibilities. They have to depend greatly on the circumstances under which the experiments are performed. The a priori credibilities can even be altered in the middle of a series of experiments. As a matter of fact, in natural science or in daily life, all conceivable hypotheses are not usually thought of at the beginning. On the contrary, as a series of experiments goes on, a scientist or layman may suddenly hit upon a new hypothesis. This means that the a priori credibility of such hypothesis is zero at the beginning of the series of experiments and suddenly takes a finite value in the middle of the series of experiments. In such a case, all the past experiments can be reconsidered in the light of the new hypothesis. This satisfies Requirement (9).

As far as Requirement (10) is concerned, Section 6 will give a detailed mathematical proof based on Bayes' theorem for the anti-ergodic tendency and the inverse H-theorem. There will also be given an estimate of the inevitable fluctuations of the inductive entropy about its "average" behavior, which satisfies the inverse H-theorem.

• *D. Some general considerations*

So far the outline of our mathematical model has been explained as if every mathematical detail had a factual meaning in every inductive process. Of course in some practical cases, the numerical values of the deductive probabilities as well as those of the credibilities are very difficult to evaluate. Even in those cases, the general nature of our approach seems to give some insight into the problem of inductive inference. Particularly interesting is the way the a priori credibilities intervene in the estimation of the *a posteriori* credibilities when the experience is finite, which is always the case with human experience. It should be emphasized once more that the term *a priori* here means "not directly determined by the observed data $\mathfrak{B}$," and does not mean "independent of all human experience". Indeed, it is very often the case that the a priori credibilities can be derived from the experience of a broader scope. For instance, if the observation concerns a particular phenomenon in the domain of pure physics, most physicists will agree to give higher a priori credibilities to those hypotheses which can be expressed in terms of a differential equation than to those hypotheses which cannot. This preference is undoubtedly a result of an inductive process based on experience in a broader domain, or at a higher level, covering a great variety of physical phenomena. Therefore, the determination of the so-called *a priori* credibilities must again be subjected to the process of gradual improvement by comparison with the experience at the higher level. This inductive determination on a higher level of the a priori credibilities of a lower level will again necessitate the use of higher level a priori credibilities. This process thus has to be continued indefinitely along a long ladder of "levels." If we push this affair further and further along this ladder, which might be infinite, it is very well possible that the hypotheses in question can be formulated only in so vague and ill-defined terms that they can only be evaluated by a quasi-esthetical criterion, such as the principle of simplicity. It should be reminded that a "higher level hypothesis" means a hypothesis interconnecting and synthesizing many lower-level hypotheses. The whole structure may thus be compared to a pyramid-like network of ladders, of which the top part can at present be described only in a rather foggy fashion.

In view of these circumstances, it is usually impossible to give a definite numerical value to the a priori credibilities, and this fact leaves room for what may be called subjective elements. Because of the property (7) mentioned above, the ultimate conclusion will be free from the subjective prejudgment. However, for a finite size of experience, these subjective elements can be as great a help as a hindrance in selecting the right hypothesis. While the credibility cannot be completely free from the subjective elements, at a finite ν, the confirmability has an objective meaning.

To illustrate some of the points of our approach, let us take an example analogous to the one discussed by N. Goodman.[4] Let us consider two hypotheses:

H_1: *Copper is electrically conductive,*

H_2: *Copper observed on or before December 31, 1960 is electrically conductive and copper observed on or after January 1, 1961 is not electrically conductive.*

These two hypotheses are probabilistically equivalent with respect to all the available data up to the present. Therefore from a purely empirical standpoint, it may be argued, they should be given an equal weight. If H_2 is logically refuted by an observation on January 1, 1961, then one can take another hypothesis H'_2 similar to H_2 but with a later date. Then, H_1 and H'_2 must be again given an equal weight. From the point of view of the present paper, if two hypotheses are probabilistically equivalent, the a priori credibility becomes a decisive factor in determining their credibilities. These a priori credibilities have to be determined by an inductive inference of higher level. For a scientist, it is perfectly natural to ask credibilities of the following hypothesis K and of its negation $\overline{K}$, although this is not the unique question to be asked in this connection.

K: *The content of a basic natural law (or a highly credible hypothesis) does not depend on a particular point of time, i.e., on a date.*

$\overline{K}$: *Negation of K.*

Goodman[4] has shown that we can rephrase H_1 and H_2 in such a way that the time-dependence appears in a statement equivalent to H_1 and the time-independence appears in a statement equivalent to H_2. But this was done only by concealing the time-dependence in a symbolic predicate. We are not interested in syntactical content of a proposition, but in the pragmatical, extrasyntactical content of propositions. Thus H_1 is pragmatically time-independent and H_2 is pragmatically time-dependent.

Now hypothesis K has a perfect match with past experience, in the sense that the deductive probability distribution (0 and 1 here) and the actual empirical probability distribution coincide. On the other hand, it is not so with hypothesis $\overline{K}$. If hypothesis $\overline{K}$ is true, then there is a nonvanishing *deductive* probability that a natural law changes on a particular date, future or past. But no basic natural law is known to have changed in the past. This means that past experience shows that the events to which the hypothesis gives nonvanishing probability have not happened. Hypothesis K is then favored, although neither hypothesis is logically refuted. This corresponds to the example quoted under Requirement (3). Then K has a perfect confirmability, while $\overline{K}$ has not, and if the a priori credibilities for K and $\overline{K}$ are not extremely different, then the a posteriori probability of K is higher than that of $\overline{K}$.

However, K and $\overline{K}$ again may be given a priori credibilities overwhelmingly in favor of $\overline{K}$. Then, this difference can offset the empirical evidence in favor of K. In fact, there may be a philosopher who believes that the hypothesis J that everything changes in time enjoys an overwhelmingly large credibility. This may cause him to attach an extremely large a priori credibility to $\overline{K}$. It is in principle possible to examine empirically the credibility of J and that of its negation $\overline{J}$, but there always remains an arbitrary element due to the a priori credibilities of J and $\overline{J}$. The philosopher can take advantage of this arbitrariness to increase the credibility of J.

The main point of this paper as applied to the present example is that for any given ratio of the a priori credibilities given to K and $\overline{K}$, the empirical data sufficiently accumulated (if in favor of K) will finally give a higher credibility to K than to $\overline{K}$. On the other hand, it is not denied that for a finite size of empirical data, an extremely high a priori credibility attached to $\overline{K}$ may offset the empirical data in favor of K. This may seem to be an endless seesaw game, but it is not quite so. The reason is that an infinitely large empirical data, if available and if in favor of K, will give, according to our theory, a credibility equal to unity to K, no matter how large an a priori credibility one may attach to $\overline{K}$. This means that the see-saw game is doomed to end in accordance with the edict of the empirical data, whatever it may be. As far as K and $\overline{K}$ are concerned, one might say that the game (at today's level of scientific knowledge) has advanced so much in favor of K that one can no longer offset it except by "distorting" the entire picture of the universe to a ridiculous degree. An example would be undue emphasis on a thesis like J, which will certainly ignore the practicalities of life. The term *distortion* as used here might be best interpreted as the a priori credibilities in flagrant contradiction to the confirmabilities determined by experience.

• *E. Intent and extent of hypothesis*

So much for the generalities of inductive inference. We shall now briefly summarize the contents of Section 2, which deals with deductive inference. We shall first define $\mathfrak{IC}$ and $\mathfrak{D}$ more closely and clarify the function of (a special type of) deductive inference. The definitions there will be made so that the relationship between the elements of $\mathfrak{IC}$ and the elements of $\mathfrak{D}$ can represent not only the relation between hypothesis and observational datum but also, to some extent, the relation between concept and particular object and the relation between pattern and individual figure. The main interest in Section 2 is the information balance in the process of deductive inference, whereby the information is defined with respect to the individual outcome of observation. The knowledge that a certain phenomenon obeys a certain law contains a certain amount I_L of information, but this is not enough to identify the individual outcome. For the latter end, one needs an additional information in the amount of I_C, which may be supplied by auxiliary conditions, such as the initial or boundary conditions. On the other hand, each law (or hypothesis) has its *extent*, E, which measures essentially how many individual outcomes are allowed by the law. Each law has also its *intent*, J, which measures essentially how specific and restrictive the law is. The results of Section 2 are expressed by three theorems, $I_L=J$, $I_C=E$, $E+J=$constant, where "constant" means "independent of which hypothesis is the law". The last relation shows that the

larger the intent the smaller the extent. The intent of a law or a concept usually means something more semantic, but what is called "intent" here may be considered to be the numerical aspect of the "intent", somewhat in a similar way as the term "information" has both semantic and numerical aspects. For instance, the *semantic intent* of a law, $d^2y/dx^2=0$, is that the curve is a straight line (among other possible curves), but the numerical intent of the same law is expressed as the logarithm of the ratio of the number of all possible curves to the number of all possible straight lines. (In order to avoid complications due to continuum, we shall interpret differentiations as differences, so as to make everything enumerable). The semantic aspect of a hypothesis is actually extremely important in selecting the hypotheses to be considered, since it can be used as a guide in retrenching $\mathfrak{H}$ from the set of all conceivable hypotheses to the set of hypotheses of a certain *type*, making the convergence of credibilities much faster. For instance, the number of hypotheses which can be expressed as $d^ry/dx^r=0$, $(r=0, 1, 2 \ldots)$, is extremely small compared with the number of all possible hypotheses, each of which represents any arbitrary class of curves. The semantic intent is also extremely important in concept building and pattern recognition. The semantic intent of a concept is the internal bondage existing among the elements and making them cohere as one family. The stronger, i.e., the more restrictive, this bondage is, the smaller the family will be; and the numerical intent measures this smallness.

- *F. Source of information*

It should be clearly understood from the foregoing that the "extent" of a law is the uncertainty (entropy) regarding the outcome of an individual observation, while the inductive entropy is the uncertainty regarding the correct hypothesis. In inductive inference, the credibility starting from a widely spread distribution over many hypotheses gradually concentrates on one hypothesis, as neatly expressed by our inverse H-theorem. Thus we are, in some sense, absorbing information regarding the correct hypothesis as we accumulate empirical data. But through this decrease of inductive entropy in inductive process, the evaluation of the extent of a law becomes more reliable, i.e., the uncertainty about the uncertainty about the outcome of an individual observation becomes smaller. Other than these two kinds of entropy functions, there is a third kind of entropy function which, as shown elsewhere,[5] increases in the process of deductive inference. All of these entropy functions represent certain kinds of information quantities, which are interrelated in a certain intricate fashion. This paper gives very little consideration to this matter, except in Section 6E.

In any event, it is a very interesting future task to investigate the flow and balance of information in all inductive and deductive activities.[6] We can, however, immediately foresee two difficulties in such an investigation. One of them concerns a question as to whether the natural phenomena in themselves (without preconceived concepts or categories) have any definite information content, which we supposedly absorb in the inductive process. The second concerns a question as to what is the real source of information contained in a very high level hypothesis, such as the principle of simplicity. Is it really to be found, as our explication may seem to lead us to believe, in experience?

As a mild warning against too hasty a mechanistic interpretation of cognitive processes, let us borrow a few words of wisdom from Gaston Bachelard[7] who enounced them in an entirely different context: "On ne peut étudier que ce qu'on a d'abord rêvé. La science se forme plutôt sur une rêverie que sur une expérience et il faut bien des expériences pour effacer les brumes du songe". The reader may be tempted to add either one of the following two rejoinders: (1) "Where there is no experience at all, there will be no reverie either," or (2) "Where there is no dreaming consciousness, there will be no experience either."

2. Basic concepts useful in describing deductive inference

- *A. Definitions*

We are given two sets of propositions, $\mathfrak{H}$ $\{H_1, H_2, \ldots, H_I, \ldots H_N\}$ and $\mathfrak{D}$ $\{D_1, D_2, \ldots, D_i, \ldots D_n\}$, and to each element H_I of $\mathfrak{H}$ are ascribed deductive conditional probabilities $p(D_i|H_I)$, or simply $p(i|I)$, $i=1, 2, \ldots, n$, such that $p(i|I)$ is the probability of proposition D_i being true when proposition H_I is true. These definitions are the mathematical basis needed in the following consideration, but some illustrations will be adduced to indicate the areas to which our mathematical formalism is meant to apply. Proposition D_i may be of the following type: By measurement of a certain physical quantity in a physical system prepared according to a given prescription, one obtains a certain value, say, V_i. Then H_I will be a hypothesis, that is, a would-be law, which is supposed to govern a certain domain of natural phenomena including the observation involved in D_i. This is a typical case of hypotheses and relevant empirical observations. We can also apply the present framework of theory, with some necessary caution, to the questions of "concepts and particulars," "patterns and figures," and "genera and species." Proposition H_I may be of the type: "The letter is A (abstract concept or pattern)," and D_i may be a proposition of the type: "The letter is found to be written as $\mathfrak{A}$ (particular item or figure)." The H's will often be referred to as hypotheses or patterns, and the D's will be referred as data or items. We assume in the present paper that N and n are finite, except in a limiting case. In some cases, one H may be a conjunction of more than one hypothesis.

In some cases, an observation may find more than one D to be true, if the D's are not disjoint (not mutually exclusive). In this case, we can replace $\mathfrak{D}$ by another $\mathfrak{D}'$ whose elements are disjoint and can be expressed as conjunctions of elements of $\mathfrak{D}$, so that the elements of $\mathfrak{D}$ can be expressed as disjunctions of elements of $\mathfrak{D}'$. If $p(i|I)$ is given with respect to $\mathfrak{D}$, and if D_i are independent, then we can calculate $p(i|I)$ with respect to $\mathfrak{D}'$ from $p(i|I)$ with respect to $\mathfrak{D}$. We shall limit ourselves to the cases where the elements of $\mathfrak{D}$ are disjoint. Thus, we assume

$$p(i|I)\geq 0; \quad \sum_{i=1}^{n} p(i|I)=1, \quad I=1, 2, \ldots, N . \qquad (2.1)$$

In some cases, particularly in problems of patterns, it happens that for a given I, $p(i|I)$ are zero for a certain group of i's, but the actual values of nonzero components of $p(i|I)$ are indeterminate. The set of i's for which $p(i|I) \neq 0$ for a given I will be called the domain of I; the number of i's in the domain will be called the dimension of I (denoted by $W(I)$). The total number of possible patterns, as defined by their domains (i.e., not by the actual values of nonzero $p(i|I)$), is 2^n-1. However it is often the case, for one reason or another, that N is less than its maximum value 2^n-1. This restriction of N often makes the problems of deduction and induction manipulable, since 2^n-1 is usually a prohibitively large number. When the domains of any two patterns of the set $\mathfrak{IC}$ do not overlap, then we speak of disjoint patterns. In this case, assuming that no i is uncovered by an H, we have

$$\sum_{I=1}^{N} W(I) = n \ . \tag{2.2}$$

When n and N are given, and further if the N numbers $W(1), W(2), \ldots, W(N)$, satisfying (2.2) are given, there are $n!/W(1)! \ldots W(N)!$ possible sets of disjoint patterns. Under similar circumstances, with a less restrictive condition that the numbers $W(1), W(2), \ldots W(N)$ are arbitrary so long as (2.2) are satisfied, the total number of possible sets of disjoint patterns will be given by

$$\sum_{k=0}^{N-1} (-1)^k (k)^N (N-k)^n \ .$$

We do *not* restrict ourselves, however, to the cases of disjoint patterns in the following.

Although we can derive various useful results in the cases where the values of the nonzero probabilities $p(i|I)$ are indeterminate, we shall often restrict our discussion in this note to the cases where all the probabilities $p(i|I)$ are given. To handle the pattern problems where $p(i|I) \neq 0$ are indeterminate, it is sometimes useful to assume

$$\left.\begin{array}{l} p(i|I) = 1/W(I), \text{ if } i \text{ belongs to the domain of } I \ , \\ p(i|I) = 0 \ , \text{ otherwise} \ . \end{array}\right\} \tag{2.3}$$

When a pattern is treated in this fashion, we shall speak of a "homogenized" pattern. We can easily set up experiments, as we shall see later in an example in Section 7, so as to meet the assumption (2.3).

In the general case, where $p(i|I)$ can take any value satisfying (2.1), the number of possible hypotheses H_I can be continuously infinite although n is finite. This is a very important fact in connection with the inherent difficulties of inductive inference. If there are $k(>1)$ hypotheses H_I which have the same probability distribution $p(i|I)$ for all i's in a subset of $\mathfrak{D}$, then we speak of a k-fold degenerate case and these hypotheses are said to be probabilistically equivalent in this subset of $\mathfrak{D}$. There can of course be more than one set of probabilistically equivalent hypotheses.

• B. *Intent and extent of hypothesis*

Deductive inference starts with an assumption that one and only one of the hypotheses, H_I, is true, i.e., it is a "law." Under this circumstance, the *ignorance* as to what outcome will be observed is given by the *extent* E of H_I, which is defined by

$$E(I) = -\sum_{i=1}^{n} p(i|I) \log p(i|I) \geq 0. \tag{2.4}$$

The maximum value of $E(I)$, with regard to I, is

$$(E(I))_{\max} = \log n \ , \tag{2.5}$$

if there is a hypothesis such that $p(i|I) = 1/n$. Equation (2.5) gives the maximum ignorance regarding the outcome which can be one of the n possibilities.

The difference between $(E(I))_{\max}$ and $E(I)$, i.e.,

$$J(I) = \log n - E(I)$$

$$= \sum_{i=1}^{n} p(i|I) \log n\, p(i|I) \geq 0 \ , \tag{2.6}$$

called the *intent* of hypothesis I, gives the amount of information regarding the outcome furnished by the knowledge that hypothesis H_I and only hypothesis H_I is true. For log n is the original ignorance without any knowledge about the hypothesis, and $E(I)$ is the remaining ignorance after cognizance is taken of the fact that H_I is the law. In this sense, $J(I)$ may be called the "predictive information content," or "predictive power" of hypothesis H_I.

In order to determine the outcome, one needs a total information in the amount log n, but the knowledge of the true hypothesis provides an information amount, $J(I)$. Hence, one needs an additional information in the amount log $n - J(I) = E(I)$ to specify the actual outcome. This information is thus the necessary auxiliary information to identify an individual, contingent outcome when the law is known. In this sense, it may be said to represent "contingent" information.

To facilitate understanding, the main points of the foregoing explanation will be repeated in the form of theorems with words instead of symbols. First we have

$$\text{(Theorem 1): Extent} + \text{Intent} = \text{constant}, \tag{2.7}$$

where "constant" means that it does not depend on I. The larger the extent, the smaller the intent. Next, we have

$$\text{(Theorem 2): Intent} = \text{predictive power of hypothesis,}$$

$$\text{(Theorem 3): Extent} = \text{necessary contingent information,} \tag{2.8}$$

where the predictive power of a hypothesis means the information provided by the hypothesis if it is true, and the contingent information of a hypothesis is the amount of information necessary to identify an observed datum D when one hypothesis is known to be true.

The reason why the words "extent" and "intent" are used will become clear if we consider the case of a "homogenized" pattern, (2.3). $E(I)$ is simply the logarithm of the number of cases included in I. We have here

$$E(I) = \log W(I) \ , \quad \text{and}$$

$$J(I) = \log n/W(I) \ . \tag{2.9}$$

In a strict deductive inference, only one hypothesis is supposed to be a law, but in a more general case, hypothesis H_I may be assigned a weight $Q(I)$ such that $Q(I) \geq 0$, and $\sum_I Q(I) = 1$. Then the probability given to the outcome i will be $\sum_I Q(I)p(i|I)$. The natural generalization of "extent" and "intent" will then be

$$E = -\sum_{i=1}^{n}\left\{\sum_{I=1}^{N}[Q(I)p(i|I)] \log\left[\sum_{I=1}^{N}Q(I)p(i|I)\right]\right\} ,$$

$$J = \log n - E , \tag{2.10}$$

which refers not to a single hypothesis but to our probabilistic knowledge about possible hypotheses. The meanings of E and J as predictive power and necessary auxiliary information or contingent information remain the same.

● *C. Example 1*

Take a rectangular coordinate system, x,y, in a plane, where x and y can take P discrete values, P being a prime number, $x,y = 0, 1, 2, \ldots, P-1$. The items considered are single-valued functions, y of x, i.e., we assume that there is one and only one value of y for each value of x. There are $n = P^P$ such curves (items), and the number of possible hypotheses is $2^{P^P}-1$. It can be shown that there is one-to-one correspondence between such a "curve" and an expression

$$y = a_{P-1}x^{P-1} + a_{P-2}x^{P-2} + \ldots + a_1x + a_0, \ (\text{mod. } P) , \tag{2.11}$$

where each of $a_0, a_1, \ldots, a_{P-1}$, has the same domain as x and y, i.e., $0, 1, 2, \ldots, P-1$. The expression (2.11) thus contains P^P different cases, as it should.

Defining "differentiation" by

$$f'(x) = \frac{\Delta f(x)}{\Delta x} = f(x+1) - f(x) , \tag{2.12}$$

where f is a function of x, we can also express any curve (2.11) by a "Taylor" series

$$\begin{aligned} y(x) = y(0) &+ \frac{y'(0)}{1!}x + \frac{y''(0)}{2!}x(x-1) + \ldots \\ &+ \frac{y^{(r)}(0)}{r!}x(x-1)\ldots(x-r+1) + \ldots \\ &+ \frac{y^{(P-1)}(0)}{(P-1)!}x(x-1)\ldots(x-P+2) \end{aligned} \tag{2.13}$$

$$= \sum_{r=0}^{P-1} y^{(r)}(0)\binom{x}{r} . \tag{2.14}$$

Here, $y^{(r)}(0)$ is the r^{th} "derivative" of y with respect to x at $x=0$, to which one is allowed to add any integral multiple of P, so that $y^{(r)}(0)$ becomes divisible by $r!$ This is always possible when P is a prime number, and the coefficient $\frac{y^{(r)}(0)}{r!}$ of $x(x-1)\ldots(x-r+1)$ becomes unique, and takes any value from $0, 1, 2, \ldots, P-1$. Thus (2.14) can be written

$$y(x) = \sum_{r=0}^{P-1} b_r x(x-1)\ldots(x-r+1) ,$$

$$b_r = 0, 1, 2, \ldots P-1 , \tag{2.15}$$

where the term for $r=0$ under the summation is understood to mean a constant b_0.

In this expression, there are P coefficients b_r, each of which can take any of P different values. Thus (2.13) or (2.15) contains $n = P^P$ items. We can thus characterize an item (curve) by a sequence of numbers $A = (a_0, a_1 \ldots, a_{P-1})$ or $B = (b_0, b_1, \ldots, a_{P-1})$.

Suppose now the set $\mathfrak{C}$ of patterns consists of P patterns defined by

$$\frac{\Delta^r y(x)}{\Delta x^r} = 0 , \quad r = 1, 2, \ldots P . \tag{2.16}$$

The solution of (2.16) can be written as

$$y(x) = \sum_{r=0}^{x-x_0} y^{(r)}(x_0)\binom{x-x_0}{r} ,$$

for $x \geq x_0$, and, in particular, for $x_0 = 0$, we have (2.14) or (2.15) with

$$B = (b_0, b_1, \ldots, b_{r-1}, 0, 0 \ldots 0) . \tag{2.17}$$

The pattern (2.16) contains P^r different items (curves). Hence, pattern $H(r)$ defined by (2.16) has a domain of dimension $W(r) = P^r$. Assuming the equal probability to each curve, i.e., assuming the homogeneity hypothesis (2.3), one obtains the extent and intent of pattern $H(r)$ from (1.9),

$$\left.\begin{aligned} E(r) &= r \log P , \\ J(r) &= (P-r) \log P , \end{aligned}\right\} \tag{2.18}$$

satisfying $E(r) + J(r) = P \log P = \log n$.

When no knowledge is available about a curve, all P coefficients in $A = (a_0, a_1, \ldots, a_{P-1})$ or in $B = (b_0, b_1, \ldots, b_{P-1})$ are arbitrary, each coefficient being capable of taking any one of P possible values. If each curve has the same probability, i.e., if we have (2.3), our ignorance about the curve is given by $\log n = P \log P$. By the knowledge that a curve belongs to $H(r)$, the number of arbitrary coefficients is reduced to r, each coefficient taking one out of P possible values. Thus the ignorance is reduced to $r \log P$. The decrease in ignorance, i.e, information, due to the knowledge of the pattern $H(r)$, is $J(r) = (P-r) \log P$, which is the predictive power of the pattern $H(r)$. When the pattern is known, then all we need in order to specify one particular curve is to determine r coefficients, $b_0, b_1, \ldots, b_{r-1}$, which allows P^r combinations. The information which allows one to select one out of P^r equally probable possibilities is $E(r) = r \log P$. Thus, this quantity is the necessary auxiliary information to specify one particular curve when it is known to belong to $H(r)$.

In deduction of a particular solution from a "natural law" expressed by a differential equation

$$\frac{\Delta^r y}{\Delta x^r} = 0 , \tag{2.19}$$

one needs the initial conditions given by r values:

$$y(0), y'(0), \ldots, y^{(r-1)}(0) \ . \tag{2.20}$$

The natural law (2.19) provides information $J(r)$ and the initial conditions (2.20) provides information $E(r)$.

We shall give an example of curves and their expressions in the forms (2.11), (2.13) and (2.15) for the case $P=7$.

Figure 1 gives the curve from which we can calculate all the derivatives, as shown in Fig. 2. The expansions (2.14) and (2.11) are in this case

$$y(x)=2+x+2x(x-1)(x-2)+2x(x-1)(x-2)(x-3)$$
$$=2+2x^2+4x^3+2x^4 \ , \quad \text{(mod. 7)} \ .$$

This can be understood as a special case of the law

$$\frac{\Delta^5 y}{\Delta x^5}=0 \ ,$$

with initial conditions

$$y(0)=2, \ y^{(1)}(0)=1, \ y^{(2)}(0)=0 \ ,$$
$$y^{(3)}(0)=5, \ y^{(4)}(0)=6 \ .$$

The law provides information in the amount $(7-5)\log 7$, while the initial conditions provide information in the amount $5 \log 7$, the sum $7 \log 7$ being just sufficient to identify one out of 7^7 possible curves.

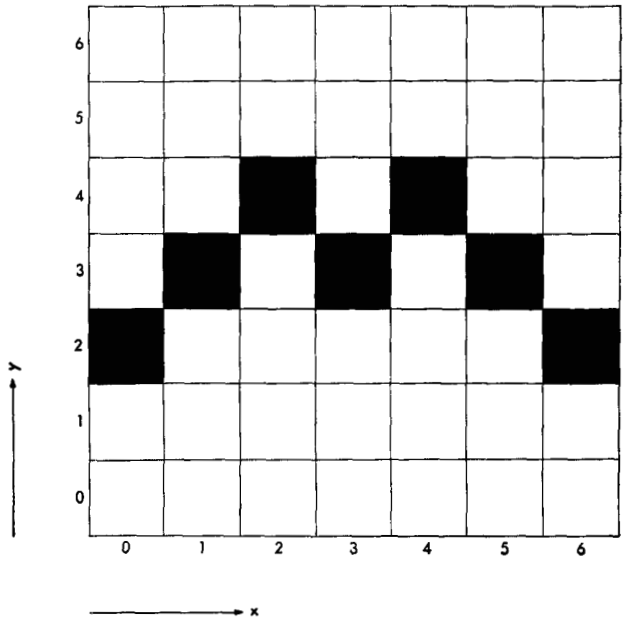


Figure 1
Example of a process of deductive inference.
The M-shaped curve here is expressed as $y=2+2x^2+4x^3+2x^4$ *(mod. 7).*

• *D. Example 2*

There are five urns, of which a certain number contain only black balls and the rest contain only white balls. The hypothesis $H(r)$ states that $r(=0, 1, 2, \ldots, 5)$ urns contain only black balls and $(5-r)$ urns contain only white balls. One is supposed to pick one ball out of an arbitrary urn without knowing which category the urn belongs to. The outcome is either black $(i=1)$ or white $(i=0)$. Thus,

$$p(1\,|\,r)=r/5 \ ,$$
$$p(0\,|\,r)=(5-r)/5 \ . \tag{2.21}$$

In this case $n=2$, $N=6$. Suppose now that $H(2)$ is known to be true, i.e., $H(2)$ is the law, then we obtain

$$E(2)=-(2/5)\log(2/5)-(3/5)\log(3/5)=0.971 \ ,$$
$$J(2)=\log 2+(2/5)\log(2/5)+(3/5)\log(3/5)=0.029 \ ,$$
$$E(2)+J(2)=\log 2=1 \ . \tag{2.22}$$

In order to obtain a unique outcome, we have to know which category an urn belongs to. The information amount which enables us to answer this last question is obviously equal to $-(2/5)\log(2/5)-(3/5)\log(3/5)$, since the probability of an urn being black is 2/5. Thus we can see that this necessary additional information is exactly equal to the "extent" $E(r)$ of (2.22).

3. Inductive probability

• *A. Credibility and Bayes' theorem as algorithm*

Inductive probability or "credibility" should be a measure of confidence we place in a hypothesis on the basis of the observed data. Credibility thus is a function of past experience, and we should not consider an a priori credibility except as an auxiliary concept, insofar as we are concerned with one particular inductive process. Our task is to explicate this vaguely conceived notion of credibility.

An experience consists of a sequence $\mathfrak{B}^{(\nu)}$ of ν consecutively observed data: $\mathfrak{B}^{(\nu)}=(i_1, i_2, \ldots, i_\nu)$, where $i_1, i_2, \ldots, i_\nu$ are the first, second, . . ., ν^{th} observed data, and can be any one of the D's defined in the previous section. On the basis of $\mathfrak{B}^{(\nu)}$ we attach to each hypothesis H_I a certain value $q(I\,|\,\mathfrak{B}^{(\nu)})$, or simply $q^{(\nu)}(I)$, in such a fashion that $q^{(\nu)}(I)$ serves "best" as an instrument to predict the result of the $(\nu+1)^{\text{th}}$ observation by the formula:

$$p^{(\nu)}(i)=\sum_I q^{(\nu)}(I)p(i\,|\,I) \ , \tag{3.1}$$

where $p^{(\nu)}(i)$ is supposed to be the probability of D_i being observed in the $(\nu+1)^{\text{th}}$ observation. Such a quantity $q^{(\nu)}(I)$ may be considered as a more explicit definition of credibility. However, there seems to be no unique way to explicate the word "best" in the foregoing sentence. The following consideration should be considered as one of the possibilities. Since $p^{(\nu)}(i)$ must be non-negative no matter what values $p(i\,|\,I)$ may have, we have from (3.1)

$$q^{(\nu)}(I)\geq 0 \ . \tag{3.2}$$

Also, by summing (3.1) over i, we obtain from (2.1)

$$\sum_I q^{(\nu)}(I)=1 \ . \tag{3.3}$$

x	y	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	$y^{(6)}$
0	2	1	0	5	6	0	0
1	3	1	5	4	6	0	
2	4	6	2	3	6		
3	3	1	5	2			
4	4	6	0				
5	3	6					
6	2						

Figure 2

Deductive inference.

The values of derivatives are derived from the values of y taken from Fig. 1. The information below the stair-shaped heavy line is not necessary if the law $\Delta^5 y/\Delta x^5 = 0$ is known.

The essential feature of inductive inference lies in the gradual change or "improvement" of $q^{(\nu)}(I)$ with ν, and there seems to be a sound reason, based on Bayes' Theorem, to assume the following formula to be an adequate expression of this process of stage-by-stage improvement, "stage" here meaning each value of ν:

$$q^{(\nu)}(I) = q^{(\nu-1)}(I)p(i_\nu|I) \Big/ \sum_{I'} q^{(\nu-1)}(I')p(i_\nu|I') \quad . \tag{3.4}$$

We shall not try to justify (3.4) but shall adopt (3.4) as an algorithm.[8]

Using (3.4) as a recurrence formula, we obtain

$$q^{(\nu)}(I) = \frac{q^{(0)}(I)p(i_1|I)p(i_2|I) \ldots . p(i_\nu|I)}{\sum_{I'} q^{(0)}(I')p(i_1|I')p(i_2|I') \ldots . p(i_\nu|I')} \quad , \tag{3.5}$$

where $q^{(0)}(I)$ may be interpreted as the a priori credibility of H_I, which cannot be determined uniquely by the observation $\mathfrak{B}^{(\nu)}$. As far as ν is finite, the $q^{(\nu)}$'s are affected by the $q^{(0)}$'s. A practically unique induction is possible only if $q^{(\nu)}(I)$ for large ν does not depend appreciably on $q^{(0)}(I)$. As far as ν is finite, one can permutate $i_1, i_2, \ldots i_\nu$ in any fashion in (3.5) without changing the values of $q^{(\nu)}(I)$. However, if $\nu \to \infty$, we have to be careful about the order of $(i_1, i_2, \ldots)$, since an arbitrary change of the order may affect the limiting behavior of $q^{(\nu)}(I)$ for $\nu \to \infty$.

It should be noted that if $q^{(0)}(I) = 0$ for a particular H_I, then $q^{(\nu)}(I) = 0$ for this H_I and for any ν. This amounts to ignoring H_I completely from our list $\mathfrak{H}$ of hypotheses from the beginning. Therefore, we assume that $q^{(0)}(I) \neq 0$ for all H_I in $\mathfrak{H}$.

In order that formula (3.5) yield an improvement of $q^{(\nu)}(I)$ rather than a meaningless fluctuation, it is necessary at least that $q^{(\nu)}(I)$ with $\nu \to \infty$ converge to a certain value

$$\lim_{\nu \to \infty} q^{(\nu)}(I) = q^{(\infty)}(I) = q(I) \quad , \quad \text{for each } I = 1, 2, \ldots, N \quad . \tag{3.6}$$

Our further interest lies in the conditions under which the limit $q(I)$, (3.6), is independent of $q^{(0)}(I)$. Such a consideration will also reveal how persistently the effect of $q^{(0)}(I)$ remains in most cases. Before undertaking to investigate this problem let us first introduce some characterizations of the observation sequence $\mathfrak{B}^{(\nu)}$.

• *B. Definition of $\mathfrak{C}^{(\nu)}$ and $\mathfrak{J}^{(\nu)}$*

$\mathfrak{B}^{(\nu)}$ is the sequence $(i_1, i_2, \ldots, i_\nu)$, in which the i's are arranged in the natural chronological order of a sequence of repeated observation. Let $\mathfrak{C}(\mathfrak{B}^{(\nu)})$, or simply $\mathfrak{C}^{(\nu)}$, be the set of the D's included in $\mathfrak{B}^{(\nu)}$. Symbolically, $\mathfrak{C}^{(\nu)} = \{D_i | D_i \in \mathfrak{B}^{(\nu)}\}$. Obviously, one has

$$\mathfrak{C}^{(1)} \subset \mathfrak{C}^{(2)} \subset \mathfrak{C}^{(3)} \subset \ldots \subset \mathfrak{C}^{(\nu)} \subset \ldots , \tag{3.7}$$

and since

$$\mathfrak{C}^{(\nu)} \subset \mathfrak{D} \text{ for any } \nu \quad , \tag{3.8}$$

the sequence of $\mathfrak{C}^{(\nu)}$ has its limit

$$\lim_{\nu \to \infty} \mathfrak{C}^{(\nu)} = \mathfrak{C}^{(\infty)} \equiv \mathfrak{C} \quad . \tag{3.9}$$

Since $\mathfrak{C}^{(\nu)}$ is a discontinuous set, $\mathfrak{C}^{(\nu)}$ will become identical with $\mathfrak{C}$ for large enough values of ν. Denoting the smallest of such ν's by ν_0, we obtain

$$\mathfrak{C}^{(\nu)} = \mathfrak{C} \quad , \quad \text{for } \nu \geq \nu_0 \quad . \tag{3.10}$$

It is sometimes necessary to consider a sequence obtained from $\mathfrak{B}^{(\infty)}$ by omitting the first $(\mu - 1)$ elements:

$$\mathfrak{B}^{(\infty)} - \mathfrak{B}^{(\mu-1)} = (i_\mu, i_{\mu+1}, \ldots). \tag{3.11}$$

The set of D's included in $\mathfrak{B}^{(\infty)} - \mathfrak{B}^{(\mu-1)}$, (3.11), will be called $\mathfrak{F}_\mu^{(\infty)}$, i.e.,

$$\mathfrak{F}_\mu^{(\infty)} = \mathfrak{C}(\mathfrak{B}^{(\infty)} - \mathfrak{B}^{(\mu-1)}); \quad \mathfrak{F}_1^{(\infty)} = \mathfrak{C} \quad . \tag{3.12}$$

We have to postulate

$$\mathfrak{F}_\mu^{(\infty)} = \mathfrak{C} \quad \text{for any finite } \mu \tag{3.13}$$

to make any intelligible discussion possible in our problem. This postulate (3.13) means essentially that if a datum D_i appears once at a finite position, it will reappear again in a position higher than any arbitrary position.

Next, let $\mathfrak{J}^{(\nu)}$ designate the set of those H's for which $p(i_\mu|I) \neq 0$ for $\mu = 1, 2, \ldots \nu$. Symbolically, $\mathfrak{J}^{(\nu)} = \{H_I | p(i|I) \neq 0$ for all $D_i \in \mathfrak{B}^{(\nu)}\}$. In other words, $\mathfrak{J}^{(\nu)}$ is the set of those I's which have nonvanishing probability $p(i|I)$ for the i's included in $\mathfrak{C}^{(\nu)}$. $\mathfrak{J}^{(\nu)}$ is obviously a function of $\mathfrak{C}^{(\nu)}$. It follows then that $\mathfrak{H} - \mathfrak{J}^{(\nu)}$ is the set of hypotheses which do not allow occurrence of $\mathfrak{B}^{(\nu)}$, i.e., those hypotheses

which should be logically refuted by $\mathfrak{B}^{(\nu)}$. From this definition, it is clear that

$$\mathcal{J}^{(1)} \supset \mathcal{J}^{(2)} \supset \cdots \supset \mathcal{J}^{(\nu)} \supset \cdots . \tag{3.14}$$

Since there is a lower bound which is the empty set, the sequence (3.14) has a limit.

$$\lim_{\nu \to \infty} \mathcal{J}^{(\nu)} = \mathcal{J}^{(\infty)} \equiv \mathcal{J} . \tag{3.15}$$

If $\mathcal{J}$ is an empty set, there will finally be no single surviving hypothesis, after logical refutation. Then the whole inductive process is worthless. Hence, we discuss only the case where

$$\mathcal{J} \neq \varnothing , \tag{3.16}$$

from which also follows, in virtue of (3.14),

$$\mathcal{J}^{(\nu)} \neq \varnothing . \tag{3.17}$$

The statement (3.16) means that there is at least one hypothesis which gives nonvanishing deductive probabilities $p(i|I)$ for all $D_i \in \mathfrak{C}$. Since $\mathcal{J}^{(\nu)}$ is a discontinuous set, $\mathcal{J}^{(\nu)}$ will become $\mathcal{J}$ for a large enough ν. Denoting the smallest of such ν by ν_0', we have

$$\mathcal{J}^{(\nu)} = \mathcal{J} \quad \text{for } \nu \geq \nu_0' . \tag{3.18}$$

We obviously have $\nu_0' \leq \nu_0$, where ν_0 is taken from (3.10) since a new member in $\mathfrak{C}^{(\nu)}$ may or may not retrench $\mathcal{J}^{(\nu)}$. Hence we still can maintain (3.18) using ν_0 for ν_0'.

If there are $k(>1)$ hypotheses H_I belonging to $\mathcal{J}^{(\nu)}$ which have the same distribution $p(i|I)$ for the i's included in $\mathfrak{C}^{(\nu)}$, we say that these hypotheses are probabilistically equivalent with respect to $\mathfrak{C}^{(\nu)}$, and we speak of a k-fold degeneracy in $\mathfrak{C}^{(\nu)}$. There can be more than one set of equivalent hypotheses with respect to $\mathfrak{C}^{(\nu)}$.

• C. *Some immediate consequences of our algorithm*

The above definition of $\mathfrak{C}^{(\nu)}$ and $\mathcal{J}^{(\nu)}$ has nothing to do with our specific algorithm (3.4) or (3.6). We shall now consider some of the consequences that can immediately be drawn from this algorithm. First, we note that the factor $p(i_1|I)p(i_2|I) \ldots . p(i_\nu|I)$ multiplying $q^{(0)}(I)$ in (3.5) is zero for a hypotheses I belonging to $\mathfrak{K} - \mathcal{J}^{(\nu)}$ and nonzero for a hypothesis belonging to $\mathcal{J}^{(\nu)}$. Since $\mathcal{J}^{(\nu)} \neq \varnothing$, (3.17), and $q^{(0)}(I) \neq 0$, the denominator of (3.5) is nonzero. If $\mathcal{J} = \varnothing$, then $q^{(\nu)}(I)$ for $\nu \geq \nu_0$ would become indeterminate. Hence, we get

$$q^{(\nu)}(I) = 0 \quad \text{for } H_I \in \mathfrak{K} - \mathcal{J}^{(\nu)} \subset \mathfrak{K} - \mathcal{J} , \tag{3.19}$$

and consequently the limit (3.6) exists and

$$q(I) = 0 \quad \text{for } H_I \in \mathfrak{K} - \mathcal{J} . \tag{3.20}$$

This means that the process of logical refutation is built in our formula (3.5). From (3.19) follows that for $\nu \geq \nu_0$ we can limit I' in (3.4) and (3.5) only to those I's which belong to $\mathcal{J}$.

From the same argument, it follows also that

$$q^{(\nu)}(I) \neq 0 \quad \text{for a finite } \nu \text{ and for } H_I \in \mathcal{J}^{(\nu)} , \tag{3.21}$$

since $q^{(0)}(I) \neq 0$ and the factor multiplying $q^{(0)}(I)$ is nonzero and the denominator in (3.5) is finite. Eqs. (3.19) and (3.21) allow us to write: $\mathcal{J}^{(\nu)} = \{H_I | q^{(\nu)}(I) \neq 0\}$ for a finite ν.

If we put $q^{(\nu)}(I) = 1$ in (3.5), it follows that the numerator and the denominator must be the same, which means that $\mathcal{J}^{(\nu)}$ consists of only one hypothesis. Consequently, at a finite ν, $q^{(\nu)}(I)$ can never become unity except in the case where all but one hypothesis has logically been refuted by then.

• D. *The limit of $q^{(\nu)}(I)$ for $\nu \to \infty$*

It is very important to note that we cannot put $\nu = \infty$ in (3.21), for, $q^{(\nu)}(I) \neq 0$ can gradually tend to zero at the limit $\nu \to \infty$. The existence of the limit (3.20) for $H_I \in \mathfrak{K} - \mathcal{J}$ follows from (3.5). But, for $H_I \in \mathcal{J}$, we have so far no guarantee that the limit (3.6) exists. We shall see in the next section a sufficient condition for the existence of the limit (3.6). In this section, however, we shall take the converse approach and assume the existence of (3.6) for $H_I \in \mathcal{J}$ and examine the consequences of this assumption. We can then classify the H_I in $\mathcal{J}$ in two classes, namely, those for which the limit $q(I) = 0$ and those for which the limit $q(I) \neq 0$. We shall denote the set of the latter hypotheses by $\mathcal{K}$. Symbolically, $\mathcal{K} = \{H_I | q(I) \neq 0\}$. Obviously $\mathcal{K} \subset \mathcal{J}$, and we have also $q(I) = 0$ for $H_I \in \mathcal{J} - \mathcal{K}$.

Now, if $q(I)$ for $H_I \in \mathcal{J}$ exists, it is necessary, according to (3.5), that for any $\epsilon_1(>0)$ there exists a $\nu_1(\geq \nu_0)$ such that

$$|q^{(\nu-1)}(I) - q^{(\nu)}(I)| = \left| q^{(\nu-1)}(I) - \frac{q^{(\nu-1)}(I)p(i_\nu|I)}{\sum_{I'} q^{(\nu-1)}(I')p(i_\nu|I')} \right| < \epsilon_1 \quad \text{for } \nu \geq \nu_1 . \tag{3.22}$$

Since this must hold for any $\nu(\geq \nu_1)$, i_ν will take all possible i's included in $\mathfrak{F}_{\nu_1}^{(\infty)}$ defined in (3.12). Due to the assumption (3.13), this means that i_ν will take all i's included in $\mathfrak{C}$. I and I' in (3.22) are those included in $\mathcal{J}$. Eq. (3.22) will obviously be satisfied if $\lim_{\nu \to \infty} q^{(\nu)}(I) = 0$. Limiting I in (3.22) to those belonging to $\mathcal{K}$, we can conclude from (3.22) that for any arbitrary $\epsilon_2(>0)$ there exists a number ν_2 such that

$$|p(i_\nu|I) - \sum_{I'} q^{(\nu-1)}(I')p(i_\nu|I')| < \epsilon_2 \text{ for } \nu \geq \nu_2 , \tag{3.23}$$

where I belongs to $\mathcal{K}$ and I' belongs to $\mathcal{J}$, and i_ν is any one of the i's belonging to $\mathfrak{C}$. Note: $p(i|I) \leq 1$. As we assume the existence of $q(I)$, we have for any arbitrary $\epsilon_3(>0)$,

$$|q^{(\nu)}(I) - q(I)| < \epsilon_3 \quad \text{for } \nu \geq \nu_3 . \tag{3.24}$$

If both ϵ_2 and ϵ_3 were zero, one would get from (3.23), (3.24),

$$p(i|I) = \sum_{I'} q(I')p(i|I') , \tag{3.25}$$

where $i \in \mathfrak{C}$, $I \in \mathcal{K}$, $I' \in \mathcal{K}$. Now the lefthand side of (3.25) depends on I, while the righthand side does not. Hence, any two I's, say I_1 and I_2, belonging to $\mathcal{K}$, must have the same distribution.

$$p(i|I_1) = p(i|I_2) , \quad I_1, I_2 \in \mathcal{K}, i \in \mathfrak{C} . \tag{3.26}$$

This means that $\mathfrak{K}$ must consist of a set of probabilistically equivalent hypotheses with respect to $\mathfrak{C}$. If the degree of degeneracy of these hypotheses is k, then k will be number of elements in $\mathfrak{K}$. Now this conclusion has been reached with the assumption that $\epsilon_2=0$, $\epsilon_3=0$. If ϵ_2, ϵ_3 are not zero, (3.26) will be replaced by

$$|p(i|I_1)-p(i|I_2)|<\epsilon_4\ ,\quad I_1, I_2\in\mathfrak{K}\ , \tag{3.27}$$

in such a way that we can make ϵ_4 arbitrarily small by taking ϵ_2 and ϵ_3 sufficiently small, i.e., by taking ν_2 and ν_3 large enough. In consequence, the conclusion (3.26) will hold in the limit.

As regards the values of $q(I)$ for I belonging to $\mathfrak{K}$, we can immediately see from (3.5) that

$$\frac{q(I_1)}{q(I_2)}=\frac{q^{(0)}(I_1)}{q^{(0)}(I_2)}\ . \tag{3.28}$$

Hence $q(I)$, for $H_I\in\mathfrak{K}$, depends on $q^{(0)}(I)$ in general. In order that $q(I)$ may not depend on $q^{(0)}(I)$, it is necessary thus that the set $\mathfrak{K}$ consist of only one hypothesis, and for this hypothesis

$$q(I)=1\ . \tag{3.29}$$

If there is no degeneracy in $\mathfrak{K}$ with respect to $\mathfrak{D}$ from the beginning, then (3.29) will always hold. However, it should be noted that even if there are degenerate hypotheses in $\mathfrak{K}$ with respect to $\mathfrak{D}$, they may drop out of $\mathfrak{K}$. The necessary condition is that the hypotheses belonging to $\mathfrak{K}$ have no degeneracy with respect to $\mathfrak{C}$.

• *E. Summary*

Summarizing the foregoing argument, we can conclude the following. The two conditions (3.13) and (3.16) being always assumed, the existence of the limit, $q(I)$, independent of the a priori probabilities $q^{(0)}(I)$, implies that $q(I)=0$ or 1.

Before we pass to the next section, it may be of interest to note that the definition of $\mathfrak{J}^{(\nu)}$, therefore of $\mathfrak{J}$, is made solely with the help of the distinction between $p(i|I)=0$ and $\neq 0$. Therefore, even without the knowledge of the precise values of the nonvanishing $p(i|I)$, we can define $\mathfrak{J}^{(\nu)}$ and $\mathfrak{J}$. This process of retrenching the set of possible hypotheses from $\mathfrak{K}$ to $\mathfrak{J}$ through the intermediary stages $\mathfrak{J}^{(\nu)}$ corresponds to a gradual elimination of inadmissible hypotheses by the counterexamples presented by the observation $\mathfrak{B}^{(\nu)}$. This is a "logical" process based on the concepts of "allowed" and "forbidden." The further retrenchment of $\mathfrak{J}$ to $\mathfrak{K}$ is possible only by a probabilistic consideration, which requires the knowledge of the values of nonvanishing $p(i|I)$. It may also be repeated that no hypothesis will be given credibility unity at a finite stage, except in the case where $\mathfrak{J}$ consists of only one element, i.e., except in the case where all but one hypothesis are logically eliminated. Also, if $q^{(0)}(I)$ are cleverly given, then $q^{(\nu)}(I)$ will approach $q(I)$ at an earlier stage of ν.

4. $\mathfrak{B}$ with definite frequency distribution

• *A. Consequences of definite frequency distribution*

In the preceding section, we assumed the existence of $q^{(\infty)}(I)$ independent of the a priori probabilities $q^{(0)}(I)$, and concluded that the values of $q^{(\infty)}(I)$ must then be zero or one. In that argument, we required a minimum property of $\mathfrak{B}^{(\nu)}$, namely $\mathfrak{F}_\mu{}^{(\infty)}=\mathfrak{C}$, (3.13). As regards $q^{(0)}(I)$, we postulated only $q^{(0)}(I)\neq 0$ for all I. In this section, we require a stronger restriction on $\mathfrak{B}^{(\nu)}$ and shall show that $q^{(\infty)}(I)$ then indeed exists in a certain sense of average. The requirement on $q^{(0)}(I)$ is again only $q^{(0)}(I)\neq 0$ for all I.

Take the empirical sequence $\mathfrak{B}^{(\nu)}=(i_1, i_2, \ldots i_\nu)$ and let ν_i be the number of times D_i appears in this sequence of length ν. We define empirical frequency distribution $\alpha_i{}^{(\nu)}(=1, 2, \ldots, n)$ by

$$\alpha_i{}^{(\nu)}=\frac{\nu_i}{\nu}\ , \tag{4.1}$$

which obviously satisfy

$$\alpha_i{}^{(\nu)}\geq 0\ ,\quad \sum_i\alpha_i{}^{(\nu)}=1\ . \tag{4.2}$$

The basic equation (3.5) can be written as

$$q^{(\nu)}(I)=\frac{q^{(0)}(I)F^{(\nu)}(I)}{\sum_{I'}q^{(0)}(I')F^{(\nu)}(I')}\ , \tag{4.3}$$

with

$$F^{(\nu)}(I)=\prod_{i=0}^{n}[p(i|I)]^{\nu_i}=\prod_{i=0}^{n}[p(i|I)]^{\alpha_i{}^{(\nu)}\nu}\ , \tag{4.4}$$

where 0^0 will be understood as meaning "one". By virtue of (3.16) and the condition $q^{(0)}(I)\neq 0$, the relations (3.2) and (3.3) are guaranteed.

We require of $\mathfrak{B}^{(\nu)}$ in this section that the appearance of outcome D_i in $\mathfrak{B}^{(\nu)}$ be governed solely by the independent probability Prob. $(D_i)=\gamma_i$. In this case, we shall say that $\mathfrak{B}^{(\nu)}$ has a *definite frequency distribution*. If we take one single $\mathfrak{B}^{(\nu)}$, the actual frequency $\alpha_i{}^{(\nu)}$ of D_i in it will not be exactly γ_i. However, if we take a very large number of samples of $\mathfrak{B}^{(\nu)}$ (with a finite ν), the average of the frequency $\alpha_i{}^{(\nu)}$ in this population will be γ_i. We shall therefore be allowed to write (4.4) as

$$F^{(\nu)}(I)=\{\prod_{i=0}^{n}[p(i|I)]^{\gamma_i}\}^{\nu} \tag{4.6}$$

to describe the "expected behavior" of $q^{(\nu)}(I)$ for a finite ν.

The deviation of $q^{(\nu)}(I)$ for large ν in an individual $\mathfrak{B}^{(\nu)}$ from its "expected" value can be estimated on the basis of the following calculation. The fluctuation δ_i of ν_i, from its expected value $\nu\gamma_i$

$$\delta_i=\nu_i-\nu\gamma_i\ ,\quad \sum_{i=1}^{n}\delta_i=0 \tag{4.7}$$

obeys the multinominal distribution, which for large ν becomes

$$\text{Prob. }(\delta_1, \delta_2, \ldots, \delta_n)=[(2\pi\nu)^{n-1}\prod_{i=1}^{n}\gamma_i]^{-\frac{1}{2}}\exp\left(-\sum_i\delta_i{}^2/2\nu\gamma_i\right)\ . \tag{4.8}$$

The expected values of $\delta_i\delta_j$, under the restriction $\sum_{i=1}^{n}\delta_i=0$,

is given by

$$\overline{\delta_i\delta_j} = -\nu\gamma_i\gamma_j \text{ for } i\neq j; \quad \overline{\delta_i^2} = \nu\gamma_i(1-\gamma_i) \ . \tag{4.9}$$

Consequently, the rms fluctuation of $\alpha_i^{(\nu)}$ from γ_i is

$$\sqrt{\overline{(\alpha_i^{(\nu)}-\gamma_i)^2}} = \sqrt{\gamma_i(1-\gamma_i)/\nu} \tag{4.10}$$

which tends to zero with $\nu\to\infty$. Thus we can write

$$\alpha_i \equiv \alpha_i^{(\infty)} = \lim_{\nu\to\infty} \alpha_i^{(\nu)} = \gamma_i \ , \quad i=1, 2, \ldots, n \ . \tag{4.11}$$

See Section 6C for the fluctuations of ν_i and $q^{(\nu)}(I)$.

We have to introduce at this stage a simple lemma. Let $G^{(\nu)}(I)$, $I=1, 2, \ldots, N$, be a probability distribution with regard to I given by

$$G^{(\nu)}(I) = \frac{[A(I)]^\nu}{\sum_{I'}[A(I')]^\nu} \ , \tag{4.12}$$

where $A(I)$, $(I=1, 2, \ldots, N)$, are non-negative and not all of them are zero. Let the largest value of $A(I)$ be denoted by

$$\operatorname*{Max}_I (A(I)) = A_{\max} \ , \tag{4.13}$$

and let m be the number of I's which have this value $A_{\max}$. Then we have the following lemma.

$\lim_{\nu\to\infty} G^{(\nu)}(I) = 1/m$ for those I's (whose number is m) such that $A(I) = A_{\max}$,

$$\lim_{\nu\to\infty} G^{(\nu)}(I) = 0 \text{ for other } I\text{'s} \ . \tag{4.14}$$

The proof is almost unnecessary, since if we have for two I's, say I_1 and I_2, $A(I_1)>A(I_2)$, then $\lim_{\nu\to\infty}[A(I_1)/A(I_2)]^\nu\to\infty$.

By writing (4.6) as

$$F^{(\nu)}(I) = [(A(I))]^\nu \ , \ A(I) = \prod_{i=0}^{n}[p(i|I)]^{\gamma_i} \ , \tag{4.15}$$

we immediately obtain from (4.3) and (4.14)

$$\lim_{\nu\to\infty} q^{(\nu)}(I) = \frac{q^{(0)}(I)}{\sum_{I'}' q^{(0)}(I')} \ , \tag{4.16}$$

for those I's for which $A(I)=A_{\max}$. The summation Σ' is also extended only over such I's. We have also

$$\lim_{\nu\to\infty} q^{(\nu)}(I) = 0 \tag{4.17}$$

for those I's for which $A(I)\neq A_{\max}$. If there are k degenerate I's for which $A(I)=A_{\max}$, then these k I's belong to the class of I's under consideration with respect to (4.16). However, conversely $A(I_1)=A(I_2)=A_{\max}$ does not necessarily imply that I_1 and I_2 are probabilistically equivalent. $\mathcal{J}$ is characterized by $A(I)\neq 0$, and $\mathcal{K}$ by $A(I)=A_{\max}$. $\mathcal{K}$-$\mathcal{J}$ and $\mathcal{J}$-$\mathcal{K}$ are respectively characterized by $A(I)=0$ and $0\neq A(I)\neq A_{\max}$.

Now if there is only one I for which $A(I)=A_{\max}$, then $q^{(\nu)}(I)$ becomes independent of $q^{(0)}(I)$ and

$$\lim_{\nu\to\infty} q^{(\nu)}(I) = 1 \text{ or } 0 \ . \tag{4.18}$$

The fact that $\mathcal{K}$ does not necessarily consist of probabilistically equivalent hypotheses may seem to be at variance with the conclusion of the last section, where $\mathcal{K}$ consisted of probabilistically equivalent hypotheses. This apparent discrepancy stems from the fact that in the last section we assumed the uniform convergence of $q(I)$ for any arbitrary $\mathfrak{B}^{(\nu)}$ obeying only (3.13), while in this section we have proved the convergence of average $q(I)$ for an "expected" $\mathfrak{B}^{(\nu)}$ obeying a particular probabilistic law given by the γ's. The convergence of $q(I)$, as assumed in the last section, is a more stringent notion than the convergence considered in this section. However, this rather delicate discussion becomes unnecessary if $\mathfrak{C}$ contains a hypothesis for which $\gamma_i = p(i|I)$. For, in this case, as we shall see later in Section 5, the relation $A(I)=A_{\max}$ is satisfied only by those hypotheses which are statistically equivalent in $\mathfrak{C}$ satisfying $\gamma_i = p(i|I)$.

Summarizing the results, we can say, under the assumption (3.16), that if the empirical sequence has a definite frequency distribution, then as far as the expected behavior is concerned the limits

$$\lim_{\nu\to\infty} q^{(\nu)}(I) = q(I) \tag{4.19}$$

exist and their values are 0 or given by (4.16). In particular, if there is only one I for which $A(I)=A_{\max}$, then the limits are either 0 or 1. See Section 6 for the fluctuations about this expected behavior in this case, in particular, (6.22) and (6.26).

• B. *Example 1*

It is known that an urn contains a very large number N of balls, of which some are white and the rest are black. There are $N+1$ possible hypotheses H_I $(I=0, 1, 2, \ldots N)$, namely, H_I: The urn contains I white balls and $N-I$ black balls.

The experiment consists in taking one ball from the urn, determining its color and replacing it back into the urn. There are thus two possible data, or events:

D_1: The ball is white,
D_2: The ball is black.

Suppose we repeated ν experiments, and obtained ν_1 times D_1 and ν_2 times D_2. The problem is to obtain the inductive probability for H_I. We have here

$$p(D_1|I) = \frac{I}{N} \ , \ p(D_2|I) = 1-\frac{I}{N} \ , \tag{4.20}$$

$$q^{(\nu)}(I) = \frac{q^{(0)}(I)\left(\frac{I}{N}\right)^{\nu_1}\left(1-\frac{I}{N}\right)^{\nu_2}}{\sum_{I'} q^{(0)}(I')\left(\frac{I'}{N}\right)^{\nu_1}\left(1-\frac{I'}{N}\right)^{\nu_2}} \ . \tag{4.21}$$

Since N is extremely large, let us put

$$\frac{I}{N} = x \ , \ q^{(\nu)}(I) = \frac{1}{N}q^{(\nu)}(x) \ , \ q^{(0)}(I) = \frac{1}{N}q^{(0)}(x) \ . \tag{4.22}$$

Then $q^{(\nu)}(x)\,dx$ represents the total sum of those $q^{(\nu)}(I)$ whose argument I lies between I and $I+N\,dx$. There are $N\,dx$ such I's. Summation over I can be replaced by an integral, so that for instance,

$$\sum_{I=0}^{N} q^{(\nu)}(I) = \int_0^1 q^{(\nu)}(I) N\, dx = \int_0^1 q^{(\nu)}(x)\, dx = 1 \ . \tag{4.23}$$

(4.21) becomes

$$q^{(\nu)}(x) = \frac{q^{(0)}(x) x^{\nu_1} (1-x)^{\nu_2}}{\int_0^1 q^{(0)}(x) x^{\nu_1} (1-x)^{\nu_2} dx} \ . \tag{4.24}$$

Now the distribution function multiplying $q^{(0)}(x)$ in (4.24) is, when normalized,

$$f(x) = \frac{(\nu_1+\nu_2+1)!}{\nu_1!\nu_2!}\, x^{\nu_1}(1-x)^{\nu_2} \tag{4.25}$$

which has its maximum at

$$x = \frac{\nu_1}{\nu_1+\nu_2} \ . \tag{4.26}$$

The average and the higher moments of (4.25) are

$$\langle x \rangle = \int_0^1 x f(x)\, dx = \frac{\nu_1+1}{\nu_1+\nu_2+2} \tag{4.27}$$

$$\langle x^r \rangle = \int_0^1 x^r f(x)\, dx = \frac{(\nu_1+\nu_2+1)!}{(\nu_1+\nu_2+r+1)!} \frac{(\nu_1+r)!}{\nu_1!} \ . \tag{4.28}$$

For large $\nu_1 \gg 1$,

$$\langle x \rangle \to \frac{\nu_1}{\nu_1+\nu_2} \ , \ \langle x^2 \rangle \to \left(\frac{\nu_1}{\nu_1+\nu_2} \right)^2 \tag{4.29}$$

showing

$$(\langle x^2 \rangle - \langle x \rangle^2) \to 0 \ . \tag{4.30}$$

This means that for a very large number of trials, $\nu = \nu_1 + \nu_2 > \nu_1 \gg 1$, the distribution $f(x)$ is sharply concentrated about the mean value $x = \alpha_i^{(\nu)} = \nu_1/(\nu_1+\nu_2)$. Therefore, if $q^{(0)}(x)$ is continuous in the vicinity of this point, we obtain from (4.24)

$$q^{(\nu)}(x) = \frac{(\nu_1+\nu_2+1)!}{\nu_1!\nu_2!}\, x^{\nu_1}(1-x)^{\nu_2} \ , \tag{4.31}$$

which is the same as (4.25), or in terms of the variable I,

$$q^{(\nu)}(I) = \frac{1}{N} \frac{(\nu_1+\nu_2+1)!}{\nu_1!\nu_2!} \left(\frac{I}{N} \right)^{\nu_1} \left(1 - \frac{I}{N} \right)^{\nu_2} \ , \tag{4.32}$$

which does not depend on $q^{(0)}(I)$ any longer.

The probability of obtaining the event D_1 in the $(\nu+1)^{\text{th}}$ observation is then, according to (3.1),

$$p^{(\nu)}(D_1) = \sum_I q^{(\nu)}(I)(D_1 | I) = \frac{\nu_1+1}{\nu_1+\nu_2+2} \ , \tag{4.33}$$

which is the same as the mean value given in (4.27).

When $\nu_2 = 0$, that is, when ν observations have consecutively given the same event D_1, then the probability of obtaining D_1 in the next observation is, according to (4.33),

$$p^{(\nu)}(D_1) = \frac{\nu+1}{\nu+2} \ , \tag{4.34}$$

which is Laplace's law of succession.

Coming back to the general case, the statement that (4.31) or (4.32) is true for large ν_1 (therefore large ν), implies already that $\alpha_1^{(\nu)} = \nu_1/(\nu_1+\nu_2)$ converges for large ν, see (4.29). If ν_1 becomes extremely large, then (4.24) will become

$$q^{(\infty)}(x) = \delta(x - \alpha_1) \ , \tag{4.35}$$

where $\delta(x-\alpha_1)$ is to be considered to be zero except in a small vicinity of width $1/N$ of α_1, in which it takes value of the order of N. Therefore, (4.32) becomes

$$q^{(\infty)}(I) = 1 \ , \tag{4.36}$$

for the particular I which gives the maximum value of

$$\alpha_1 \log \frac{I}{N} + (1-\alpha_1) \log \left(1 - \frac{I}{N} \right) \tag{4.37}$$

provided there is only one such I. And for the remaining I's, we shall have

$$q^{(\infty)}(I) = 0 \ . \tag{4.38}$$

• C. *Example 2*

The following problem is an example that could be adduced as an argument against the use of the Bayes Theorem in inductive inference.[9] The author, on the other hand, would like to present here his defense of the use of the Bayes Theorem. The experiment consists in determining the "head" or "tail" of a coin after tossing it. The experiment is repeated with the same coin. The $\mathfrak{D}$ consists of "head(0)" and "tail(1)". The hypotheses in question are two:

H(F): The coin is a falsified, double-headed one.

H(G): The coin is a genuine one.

Thus:

$$p(0|F) = 1, \ p(1|F) = 0, \ p(0|G) = p(1|G) = 1/2 \ . \tag{4.39}$$

Suppose we tossed ν times and have gotten heads all ν times. What is the minimum value of ν for us to be reasonably convinced that $H(F)$ is true? To make the problem more concrete, what is the minimum value of ν to make $q^{(\nu)}(F)$ larger than 10 times as large as $q^{(\nu)}(G)$? Now the argument against the Bayes Theorem runs somewhat as follows. The a priori probability (taken as a statistical frequency in the real population of all coins genuine and falsified) for a coin to be double-headed is extremely small, say, 10^{-20}, i.e., $q^{(0)}(F)/q^{(0)}(G) = 10^{-20}$. Therefore the solution for ν of inequality

$$\frac{q^{(\nu)}(F)}{q^{(\nu)}(G)} \geq 10 \tag{4.40}$$

with

$$\frac{q^{(\nu)}(F)}{q^{(\nu)}(G)} = \frac{q^{(0)}(F)}{q^{(0)}(G)(1/2)^{\nu}} = 10^{-20} \cdot 2^{\nu} \tag{4.41}$$

is

$$\nu \geq 70 \ . \tag{4.42}$$

Now, the opponent of the Bayes Theorem says that any reasonable man will be convinced of $H(F)$ after a small ν, say, 6 or 7 at most. The Bayes Theorem is therefore unreasonably prudential.

However, this argument is based on a wrong interpretation of the a priori credibility, which depends greatly on the entire circumstances under which the experiment is made. If, for instance, one goes to a nearby grocery store and takes a penny from an arbitrary cash register, then the a priori credibility for this penny to be double-headed is extremely small, as a result of which it will take indeed a very large ν for a reasonable man to decide that the coin is false. But if a professor of psychology comes to a subject of his experiment, to test human behavior in inductive inference, then there is a great deal of probability that the professor uses a variety of tricks and gimmicks. Therefore, the values of $q^{(0)}(F)/q^{(0)}(G)$ cannot be too small. The Bayes Theorem gives $\nu \geq 7$ if $q^{(0)}(F)/q^{(0)}(G) = 1/10$, which may not be too far from a realistic situation.

The subject, of course, does not use a mathematical formula to make his decision. However, we can attempt to translate his mental process more or less in terms of mathematical formulae and give some interpretation. An outstanding feature of the actual situation is that the a priori credibility does not have a fixed value. As a matter of fact, if a psychology professor starts this experiment before a subject, the latter at the beginning may even not think of such a possibility as a double-headed coin. That means the a priori credibility $q^{(0)}(F)$ at the beginning is practically zero. But suddenly the possibility of such a hypothesis may occur to the subject, and thus $q^{(0)}(F)$ jumps from zero to a finite value. And further, guessing the motivation of the professor, the subject may still increase the value of $q^{(0)}(F)$ during the course of the experiment. This kind of consideration has nothing to do with the gradual change of $q^{(\nu)}(F)$ with the accumulation of the empirical data $\mathfrak{B}$, and therefore must be attributed to $q^{(0)}(F)$. If the coin is taken arbitrarily from a grocery store cash register, the subject would not increase $q^{(0)}(F)$ so much as with the coin taken out of the psychology professor's pocket.

An interesting fact here is that the a priori credibility can change, even in the middle of the experiments. A similar argument has previously been used by the author to repudiate Loschmidt's objection to the H-theorem.[10]

5. Confirmability

A. *Comparison of predicted frequency and observed frequency*

Equation (4.3) shows that $q^{(\nu)}(I)$ is equal to $q^{(0)}(I)$ times an empirical weight proportional to $F^{(\nu)}(I)$. The larger the $F^{(\nu)}(I)$ is, the more the hypothesis H_I is confirmed by experiments. For this reason, we may use

$$\frac{1}{\nu} \log F^{(\nu)}(I) = \log \prod_{i=0}^{n} [p(i|I)]^{\alpha_i^{(\nu)}} \tag{5.1}$$

$$= \sum_{i=1}^{n} \alpha_i^{(\nu)} \log p(i|I) \quad (\leq 0) \ , \tag{5.2}$$

as a measure of confirmation (per observation) of hypothesis H_I by experiments up to the ν^{th} observation.

The quantity (5.2) takes its maximum value

$$\sum_{i=1}^{n} \alpha_i^{(\nu)} \log \alpha_i^{(\nu)} \quad (\leq 0) \ , \tag{5.3}$$

if there is a hypothesis which gives

$$p(i|I) = \alpha_i^{(\nu)} \text{ for all } i \ . \tag{5.4}$$

Therefore, the quantity

$$C^{(\nu)}(I) = \frac{\sum_{i=1}^{n} \alpha_i^{(\nu)} \log \alpha_i^{(\nu)}}{\sum_{i=1}^{n} \alpha_i^{(\nu)} \log p(i|I)} \tag{5.5}$$

is a convenient measure of confirmation, and will be called confirmability of hypotheses I at stage ν. Indeed, we have

$$0 \leq C^{(\nu)}(I) \leq 1 \ , \tag{5.6}$$

and $C^{(\nu)}(I) = 1$ if and only if the "perfect match" (5.4) takes place. Further, one gets $C^{(\nu)}(I) = 0$, if and only if there occurs an observed datum D_i which is prohibited by hypothesis H_I, i.e., there is an i for which $\alpha_i^{(\nu)} \neq 0$ and $p(i|I) = 0$. In other words, $C^{(\nu)}(I) = 0$ for a logically excluded hypothesis.

If $\mathfrak{B}$ has a definite frequency distribution as expressed by (4.11), then $C^{(\nu)}$ will converge to

$$C(I) \equiv C^{(\infty)}(I) = \frac{\sum_{i=1}^{n} \alpha_i \log \alpha_i}{\sum_{i=1}^{n} \alpha_i \log p(i|I)} \ . \tag{5.7}$$

It is of importance to note that if there is one and only one I which makes a perfect match, (5.4) with $\nu = \infty$, then that H_I will finally win out with $q(I) = 1$, since such H_I will certainly make $A(I)$, (4.15), maximum. However, the hypothesis H_I which is given credibility unity as $\nu \to \infty$ does not necessarily exhibit a perfect matching. If two hypotheses are statistically equivalent in $\mathfrak{C}$ and show both perfect matching, then their final credibilities will be proportional to their respective a priori credibilities; see (3.28).

Another quantity which may be used for the same purpose as $C^{(\nu)}(I)$ is

$$C'^{(\nu)}(I) = \frac{\sum_{i=1}^{n} p(i|I) \log p(i|I)}{\sum_{i=1}^{n} \alpha_i^{(\nu)} \log p(i|I)} \ . \tag{5.8}$$

6. Inverse H-theorem for inductive entropy

A. *Statement of theorems*

By the use of inductive probability distribution $q^{(\nu)}(I)$, one can define an "entropy" function

$$U^{(\nu)} = -\sum_{I} q^{(\nu)}(I) \log q^{(\nu)}(I) \ , \tag{6.1}$$

which measures the uncertainty about the hypotheses. The largest possible value of $U^{(\nu)}$ is log N which corresponds to the case where $q^{(\nu)}(I)$ assigns an equal probability to all hypotheses in $\mathfrak{K}$. Its smallest value is zero, which means that one of the hypotheses is true and all the rest are false. We can prove the following two theorems, which in a sense express a tendency of the $U^{(\nu)}$-function which is opposite to the H-theorem.

Theorem: If $q^{(\nu)}(I)$ has a limit for $\nu \to \infty$ independent of $q^{(0)}(I)$, and if the logical refutation leaves more than one hypothesis, then

$$U^{(\nu)} > U^{(\infty)} = 0, \ \nu = \text{finite} \ . \tag{6.2}$$

Theorem: If the empirical data has a definite frequency distribution, then except for a finite number of the first values of ν, $U^{(\nu)}$ is "expected" to decrease monotonously, i.e.,

$$U^{(\mu)} \leq U^{(\nu)}, \text{ for } \nu < \mu \ . \tag{6.3}$$

The first theorem is obvious, since $U^{(\nu)} > 0$ because of the remark in the last paragraph of Section 3 C and $U^{(\infty)} = 0$ due to (3.29). The term "expected" in the second theorem is used in the sense of (4.6), hence (6.3) represents an "average" behavior of $U^{(\nu)}$ in many series of experiments.

In an individual series $\mathfrak{B}^{(\nu)}$ of experiments, the $U^{(\nu)}$ will fluctuate about the monotonously decreasing curve, because of the effect discussed in (4.9), (4.10). Later in this section an estimation of this fluctuation for large ν will be made for the case $U^{(\infty)} = 0$ and shown to decrease with ν.

• B. *Proof of second theorem*

The premise of the second theorem is that $q^{(\nu)}$ is given, as in (4.6), by

$$q^{(\nu)}(I) = \frac{q^{(0)}(I) F^{(\nu)}(I)}{\sum_{I'} q^{(0)}(I') F^{(\nu)}(I')} \tag{6.4}$$

with

$$F^{(\nu)}(I) = [A(I)]^{\nu} \ , \tag{6.5}$$

$$A(I) = \prod_{i=0}^{n} [p(i \mid I)]^{\alpha_i} \ , \ 0 \leq A(I) \leq 1 \ . \tag{6.6}$$

In these expressions, ν can be considered as a continuous variable, and $-dU^{(\nu)}/d\nu$ will represent the information gain per stage regarding the right hypothesis. We shall demonstrate that for ν larger than a certain lower bound,

$$\frac{dU^{(\nu)}}{d\nu} \leq 0 \ , \tag{6.7}$$

from which (6.3) follows. It should be noted that the hypotheses which do not belong to $\mathfrak{J}$ have $A(I) = 0$, hence $q^{(\nu)}(I) = 0$ for $\nu \geq 1$ according to the present approximation.

Taking any one I_0 of the hypotheses belonging to $\mathfrak{K}$

$$H_{I_0} \in \mathfrak{K} \ , \tag{6.8}$$

as standard, we introduce new variables $\beta(I)$ and $x(I)$ by

$$\beta(I) = \frac{q^{(0)}(I)}{q^{(0)}(I_0)} > 0, \ 1 \geq x(I) = \frac{A(I)}{A(I_0)} > 0, \ I \in \mathfrak{J} \ . \tag{6.9}$$

$\beta(I)$ is a measure of the a priori credibility and $x(I)$ is a measure of the confirmability. $x(I) = 1$ for I belonging to $\mathfrak{K}$. Then, it is easy to see that we can write

$$U^{(\nu)} = \frac{Y^{(\nu)}}{X^{(\nu)}} \ , \tag{6.10}$$

with

$$X^{(\nu)} = \sum_{I} \beta(I) x^{\nu}(I) \tag{6.11}$$

$$Y^{(\nu)} = \sum_{I} \beta(I) x^{\nu}(I) \log \sum_{I'} \beta(I') x^{\nu}(I') - \sum_{I} \beta(I) x^{\nu}(I) \log \beta(I) x^{\nu}(I) \ , \tag{6.12}$$

where the summation with regard to I extends over $\mathfrak{J}$. Then, the derivative:

$$\frac{dU^{(\nu)}}{d\nu} = [X^{(\nu)}(dY^{(\nu)}/d\nu) - (dX^{(\nu)}/d\nu) Y^{(\nu)}]/(X^{(\nu)})^2 \tag{6.13}$$

becomes

$$\frac{dU^{(\nu)}}{d\nu} = 1/(X^{(\nu)})^2 \sum_{I} \sum_{I'} \beta(I) x^{\nu}(I) \beta(I') x^{\nu}(I') \log [\beta(I) x^{\nu}(I)] \ln [x(I')/x(I)] \ . \tag{6.14}$$

Adding to this expression another expression obtained from (6.14) by an interchange of I and I', one obtains

$$2(X^{(\nu)})^2 \frac{dU^{(\nu)}}{d\nu} = \sum_{I} \sum_{I'} \beta(I) x^{\nu}(I) \beta(I') x^{\nu}(I') \times \log \frac{\beta(I) x^{\nu}(I)}{\beta(I') x^{\nu}(I')} \ln \frac{x(I')}{x(I)} \ . \tag{6.15}$$

Now if log $[\beta(I)/\beta(I')]$ and log $[x(I)/x(I')]$ have the same sign, then we have obviously

$$\log \frac{\beta(I) x^{\nu}(I)}{\beta(I') x^{\nu}(I')} \ln \frac{x(I')}{x(I)} < 0 \ . \tag{6.16}$$

If log $[\beta(I)/\beta(I')]$ and log $[x(I)/x(I')]$ are nonzero and have opposite signs, then we have again (6.16) for ν satisfying

$$\nu > - \log \frac{\beta(I)}{\beta(I')} \Big/ \ln \frac{x(I)}{x(I')} \ . \tag{6.17}$$

If log $[\beta(I)/\beta(I')] = 0$, and log $[x(I)/x(I')] \neq 0$, then (6.16) holds for any ν. If log $[x(I)/x(I')] = 0$, then

$$\log \frac{\beta(I) x^{\nu}(I)}{\beta(I') x^{\nu}(I')} \ln \frac{x(I')}{x(I)} = 0 \tag{6.18}$$

no matter what value log $[\beta(I)/\beta(I')]$ may have. Therefore, we conclude from (6.15), (6.16), (6.18)

$$\frac{dU^{(\nu)}}{d\nu} \leq 0 \tag{6.19}$$

for ν large enough so that (6.17) is satisfied for those pairs (I, I') which have different x's. For a finite ν larger than the lower bound set by this consideration, relation (6.19) can be interpreted as the "expected" behavior of $U^{(\nu)}$. Thus the inverse H-theorem is proved. Further, if all the hy-

potheses are given equal a priori credibilities, then the lower bound of determined by (6.17) becomes zero. Hence, the theorem is true for any ν.

● C. *Evaluation of fluctuations*

If we use a single experimental series $\mathfrak{B}^{(\nu)}$ to calculate $U^{(\nu)}$ with the help of the empirical α's, then the value of $U^{(\nu)}$ is bound to show some fluctuation about the "expected" curve of $U^{(\nu)}$. However, since the fluctuation of the empirical α's, as was shown in (4.10), decreases with increasing ν, it can be expected, at least within a certain limitation, that the fluctuation of $U^{(\nu)}$ about the smooth curve also becomes very small for large ν. As an example, we shall now give an estimate of the order of magnitude of the fluctuation of $U^{(\nu)}$ for large ν in the case where $U^{(\nu)} \to 0$ as $\nu \to \infty$.

This last condition means that in the expression of $U^{(\nu)}$, (6.1), one of the $q^{(\nu)}(I)$, say, $q^{(\nu)}(I_0)$ becomes very close to unity for larger ν, and all the remaining $q^{(\nu)}(I)$ become close to zero. Because of the nature of the function $x \log x$, the contribution to $U^{(\nu)}$ from $q^{(\nu)}(I_0)$ then becomes negligible compared with the contributions from other $q^{(\nu)}(I)$, $I \neq I_0$. In the same way, in the expression of the small fluctuation $\delta U^{(\nu)}$ of $U^{(\nu)}$:

$$\delta U^{(\nu)} = -\sum_I \delta q^{(\nu)}(I) \log q^{(\nu)}(I) \ , \tag{6.20}$$

we can ignore the term corresponding to I_0. In the denominator of the expression, (4.3), of $q^{(\nu)}(I)$, the term corresponding to $I' = I_0$ will be very large compared with the other terms. Therefore, we can write (4.3) as

$$q^{(\nu)}(I) \approx \frac{q^{(0)}(I)}{q^{(0)}(I_0)} \prod_{i=1}^{n} \left[\frac{p(i|I)}{p(i|I_0)} \right]^{\nu_i}$$

$$= \frac{q^{(0)}(I)}{q^{(0)}(I_0)} \exp \left[\sum_i \nu_i \ln \frac{p(i|I)}{p(i|I_0)} \right] \quad \text{for } I \neq I_0 \ , \tag{6.21}$$

from which follows

$$\frac{\delta q^{(\nu)}(I)}{q^{(\nu)}(I)} \approx \sum_i \delta \nu_i C_i(I) \ , \tag{6.22}$$

with

$$C_i(I) = \ln p(i|I) - \ln p(i|I_0) \ . \tag{6.23}$$

With the aid of (4.9), we derive from (6.22)

$$\frac{\overline{\delta q^{(\nu)}(I) \delta q^{(\nu)}(I')}}{q^{(\nu)}(I) q^{(\nu)}(I')} \approx \nu D(I, I') \ , \tag{6.24}$$

with

$$D(I, I') = \sum_i \gamma_i C_i(I) C_i(I') - \left(\sum_i \gamma_i C_i(I) \right) \left(\sum_i \gamma_i C_i(I') \right) \ . \tag{6.25}$$

On the other hand, (6.21) shows that the order of magnitude of $q^{(\nu)}(I)$ can be written (putting $\nu_i = \nu \gamma_i$ and assuming $q^{(0)}(I) \approx q^{(0)}(I_0)$)

$$q^{(\nu)}(I) \approx e^{-\nu G(I)} \ , \tag{6.26}$$

with

$$G(I) = -\sum_i \gamma_i C_i(I) \ . \tag{6.27}$$

where $G(I)$ must be positive because $q^{(\nu)}(I)$ with $I \neq I_0$ becomes very small for large ν. From (6.24) and (6.25), we obtain

$$\overline{\delta q^{(\nu)}(I) \delta q^{(\nu)}(I')} \approx D(I, I') \nu \exp \{ -\nu [G(I) + G(I')] \} \ . \tag{6.28}$$

Substituting (6.26) and (6.28) in the square of (6.20), we obtain

$$\overline{(\delta U^{(\nu)})^2} \approx \sum_{I \neq I_0} \sum_{I' \neq I_0} D(I, I') G(I) G(I') \nu^3 \exp \{ -\nu [G(I) + G(I')] \} \ . \tag{6.29}$$

The square of the entropy $U^{(\nu)}$ itself is, according to (6.26),

$$(U^{(\nu)})^2 \approx \sum_{I \neq I_0} \sum_{I' \neq I_0} G(I) G(I') \nu^2 \exp \{ -\nu [G(I) + G(I')] \} \ . \tag{6.30}$$

Therefore we see that both $U^{(\nu)}$ and its fluctuation tends to zero with increasing ν, as was expected. It is true that $\overline{(\delta U^{(\nu)})^2}$ vanishes more slowly with increasing ν than $(U^{(\nu)})^2$, but this should not be disconcerting, since the inductive entropy $U^{(\nu)}$, in the same way as the physical entropy, does not change its usefulness if one adds to it an arbitrary constant. For instance, we can use $V^{(\nu)} = (\log N - U^{(\nu)})/\log N$ as a convenient measure of "certainty". Then the fluctuation of $V^{(\nu)}$ is very small compared with $V^{(\nu)}$ itself for large ν.

The above discussion refers to the case where $\mathfrak{K}$ consists of only one hypothesis I_0. If $\mathfrak{K}$ consists of more than one hypothesis, the situation is not so simple.

● D. *Decrease versus non-increase of entropy*

Next, we can ask whether $U^{(\nu)}$ can reach its minimum value at a finite ν. This will happen if and only if there is a number ν_1 such that $dU^{(\nu)}/d\nu = 0$ for $\nu \geq \nu_1$. To investigate this question, one should first note that the factor $\beta(I) x^{\nu}(I) \beta(I') x^{\nu}(I')$ in (6.15) is always nonzero at a finite ν. Therefore, $dU^{(\nu)}/d\nu$ becomes zero only if (6.18) is the case for all pairs (I, I') in $\mathfrak{J}$. For $I = I'$, (6.18) is self-evident. For $I \neq I'$, (6.18) holds if $x(I) = x(I')$ and/or

$$\beta(I)/\beta(I') = [x(I')/x(I)]^{\nu} \ .$$

This last equation, for a pair $x(I) \neq x(I')$, may happen to hold for a particular value of ν, but then it will not hold any value of ν larger than this value. Therefore, in order that $dU^{(\nu)}/d\nu = 0$ may hold for $\nu \geq \nu_1$, it is necessary and sufficient that $x(I) = x(I')$ for all pairs in $\mathfrak{J}$. However, as we have seen before, the I's belonging to $\mathfrak{K}$ have the largest value of the x's. Hence, the condition is satisfied only if $\mathfrak{J} = \mathfrak{K}$. If this is not the case, then $dU^{(\nu)}/d\nu$ will tend to zero only with $\nu \to \infty$. At this limit, $x^{\nu}(I)$ in (6.15) for I not belonging to $\mathfrak{K}$ will vanish because $x(I) < 1$, and the terms corresponding to two I's both belonging to $\mathfrak{K}$ will vanish because then $\ln [x(I')/x(I)] = 0$. Hence, $U^{(\nu)}$ reaches its minimum value only at the limit $\nu \to \infty$, except in a special case $\mathfrak{K} = \mathfrak{J}$. In this last case, $dU^{(\nu)}/d\nu = 0$ for any

$\nu \geq 1$. Of course, this conclusion, as well as other results discussed here, is based on (4.6).

• *E. Uncertainty about law and uncertainty about outcome*

It should be noted that

$$-\frac{dU^{(\nu)}}{d\nu}(\geq 0) \tag{6.31}$$

represents the information gain per stage regarding the limiting hypothesis, i.e., the law.

The uncertainty or ignorance $U^{(\nu)}$ regarding hypotheses should not be confused with the uncertainty regarding the outcome of an individual observation as estimated at stage ν. This last quantity should be expressed by

$$E^{(\nu)} = -\sum_i p^{(\nu)}(i) \log p^{(\nu)}(i) \ , \tag{6.32}$$

with

$$p^{(\nu)}(i) = \sum_I q^{(\nu)}(I) p(i|I) \tag{6.33}$$

in accordance with (2.10) and (3.1). A similar but different quantity is

$$<E(I)>^{(\nu)} = -\sum_I q^{(\nu)}(I) \sum_i p(i|I) \log p(i|I) \ , \tag{6.34}$$

which may be characterized as the expected value of the extent of the law. It is obvious that both quantities converge to the same value with $\nu \to \infty$, namely the extent of the limiting hypothesis. If the hypotheses are disjoint in the sense that for a given i, there is only one I such that $p(i|I) \neq 0$, then

$$E^{(\nu)} = U^{(\nu)} + <E(I)>^{(\nu)} \ , \tag{6.35}$$

but (6.35) does not hold in general cases.

The largest possible value of (6.32) is $\log n$. If the a priori probability $q^{(0)}(I)$ is such that $p^{(0)}(i) = 1/n$, then $E^{(0)}$ will be $\log n$. In such a case, we have

$$E^{(\nu)} \leq E^{(0)} \ , \quad \nu = 1, 2, \ldots . \tag{6.36}$$

This is, in a sense, comparable to the result (6.2). However, in a general case one cannot expect any definite tendency in the behavior of $E^{(\nu)}$ as a function of ν. An expression corresponding to (6.15) becomes in this case

$$2(X^{(\nu)})^2 \cdot \frac{dE^{(\nu)}}{d\nu} = \sum_i \sum_I \sum_{i'} \sum_{I'} \beta(I)\beta(I') x^{\nu}(I) x^{\nu}(I') p(i|I) p(i'|I') \times \ln \frac{x(I)}{x(I')} \log \frac{\sum_{I''} \beta(I'') x^{\nu}(I'') p(i'|I'')}{\sum_{I'''} \beta(I''') x^{\nu}(I''') p(i|I''')} \ . \tag{6.37}$$

The fact that $E^{(\nu)}$ does not show a definite tendency with regard to ν in a general case may seem disconcerting, for as we accumulate observational data one must become more "reliable" with regard to the outcome of the future observation. This disconcerting impression stems from a confusion between the concepts of "uncertainty" and "confirmability." "Uncertainty" here merely measures the statistical spread in $\mathfrak{D}$, and $E^{(\nu)}$ is a kind of average of this uncertainty, whereby the averaging is made with the help of $q^{(\nu)}(I)$, which at a finite ν still strongly depends on the a priori estimation $q^{(0)}(I)$.

A possible measure for the empirically accumulated information regarding the outcome of an observation may be obtained by a quantity like

$$\sum_I C^{(\nu)}(I) q^{(\nu)}(I) \left[\log n + \sum_i p(i|I) \log p(i|I)\right] \ , \tag{6.38}$$

but no simple theoretical foundation can be put forward to justify a formula of this kind.

7. Simulated experiments on IBM 704

• *A. First urn problem*

The urn contains ten balls, of which a certain fraction $I_0/10$ are white and the remaining fraction $(10-I_0)/10$ are black. The observation consists of taking one ball from the urn, determining its color and returning it to the urn ($n=2$). The considered hypotheses are eleven in number ($N=11$), and are of the type:

$H(I)$: *I balls out of the ten are white, and the remaining* $(10-I)$ *balls are black,* $I = 0, 1, 2, \ldots, 10$.

$H(I_0)$ is the correct hypothesis, i.e., the law to be discovered. The process under investigation is one in which the credibilities of hypotheses gradually concentrate on $H(I_0)$ as the number of observations increases. If we assign 0 and 1, respectively, to white and black, then we have

$$p(0|I) = I/10, \ p(1|I) = (10-I)/10 \ . \tag{7.1}$$

The machine simulation consists of producing the numbers 0 and 1 randomly at the ratio of $I_0/(10-I_0)$. For this purpose, a well-tested random number producing program has been utilized. The assumption of "definite distribution" is thus secured.

The first series of experiments was carried out with $I_0 = 3$. Under this condition, $H(0)$ and $H(10)$ will be "logically" refuted sooner or later, since $H(0)$ contradicts the appearance of one white ball and $H(10)$ contradicts the appearance of one black ball. Thus $\mathfrak{H}$ consists of $H(I)$ with $I = 0, 1, 2, \ldots, 10$ and $\mathfrak{J}$ consists of $H(I)$ with $I = 1, 2, \ldots, 9$. As regards the a priori credibilities, we have tried three different cases: (*i*) equal a priori credibilities are given to all eleven hypotheses; (*ii*) deliberately, a higher a priori credibility is given to a wrong hypothesis $H(7)$, viz., $q^{(0)}(7) = 12/22$ and $q^{(0)}(I) = 1/22$ for $I \neq 7$; (*iii*) a higher a priori credibility is given to the right hypothesis, viz., $q^{(0)}(3) = 12/22$ and $q^{(0)}(I) = 1/22$ for $I \neq 3$. The experiments were continued in each case until the observation number ν became 500 or more. The same sequence of random numbers was used in all three cases. In this sequence, $\mathfrak{H}$ was retrenched to $\mathfrak{J}$ with $\nu = 5$.

The smallest number ν_0 for which the inductive entropy satisfies the condition,

$$U^{(\nu)} < 0.01 \quad \text{for } \nu \geq \nu_0 \ , \tag{7.2}$$

was found to be $\nu_0 = 330$ in case (*i*), $\nu_0 = 330$ in case (*ii*) and $\nu_0 = 258$ in case (*iii*). The smallest number ν_1 for which the credibility for the correct hypothesis satisfies the condition,

$$q^{(\nu)}(3) > 0.99 \quad \text{for } \nu \geq \nu_1 \,, \tag{7.3}$$

was found to be $\nu_1 = 258$ in (*i*), $\nu_1 = 258$ in (*ii*) and $\nu = 115$ in (*iii*). Outside these slight numerical differences, the behavior of the different quantities was the same in all three cases.

The random sequence up to $\nu = 500$ contained 149 whites (0's) and 351 blacks (1's), the ratio of the whites to the total number being 0.298 instead of 0.3. The confirmability $C^{(\nu)}(3)$ of $H(3)$ serves also as a measure of this ratio in this case and gave 0.999984 at $\nu = 500$.

Figs. 3, 4 and 5 refer to the experiment of case (ii), i.e., the case where a high a priori credibility is given to a wrong hypothesis $H(7)$. Figure 3 describes the behavior of the credibilities for the correct hypothesis $H(3)$, for the wrong hypothesis $H(7)$ on which a high weight has initially been placed, and for the hypothesis $H(2)$ which is an immediate neighbor of the correct hypothesis. Since $p(i|2)$ is numerically close to $p(i|3)$, the chance is high that $q^{(\nu)}(2)$ remains relatively large compared with other hypotheses whose I is more removed from 3. $q^{(\nu)}(3)$, starting from $1/22 = 0.04545$ at $\nu = 0$, becomes definitely larger than 0.9 from $\nu = 203$ on. $q^{(\nu)}(7)$, starting from $6/11 = 0.54545$ at $\nu = 0$ becomes definitively less than 0.1 from $\nu = 6$ on. The quantity $q^{(\nu)}(2)$, starting from 0.04545 at $\nu = 0$, rises to higher values, including a maximum value 0.57003 at $\nu = 75$, but decreases finally to become definitively below 0.1 after $\nu = 202$. Roughly speaking, after $\nu = 200$, everything smoothly settles down towards the limiting situation. At $\nu = 500$, $q^{(\nu)}(2) = 0.131090 \times 10^{-5}$, $q^{(\nu)}(3) = 0.999986$, $q^{(\nu)}(7) < 10^{-30}$.

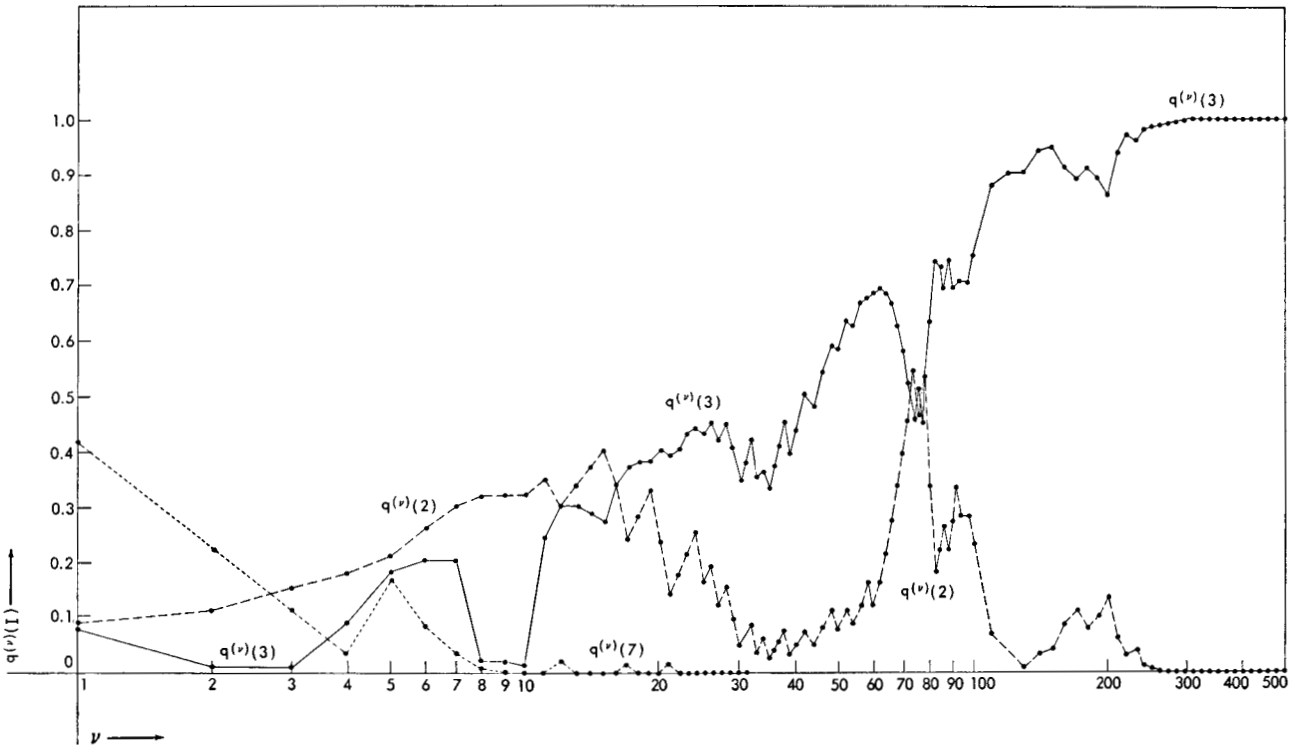
In Fig. 4 the full lines show $q^{(\nu)}(I)$ for all eleven values of I for $\nu = 0$, $\nu = 8$, $\nu = 32$ and $\nu = 128$. At $\nu = 0$, an outstanding weight is placed on $I = 7$. This effect still remains slightly at $\nu = 8$, but the weight is already shifted towards the smaller values of I. At $\nu = 32$, the entire weight is concentrated in the region of $I = 2$, 3, 4 and 5. At $\nu = 128$, $I = 3$ is already very large compared with others, although $I = 2$ and $I = 4$ still survive. For larger values of ν, of course, $I = 3$ becomes overwhelmingly large at the expense of all others. The broken line shows the confirmability $C^{(\nu)}(I)$, $I = 0, 1, 2, \ldots, 10$ at $\nu = 128$. It is 0.999962 for $I = 3$.

Fig. 5 shows the inductive entropy $U^{(\nu)}$ as function of ν. The curve, of course, is the result for one particular random sequence. If one took an average of many such sequences, one would obtain a smoothly decreasing curve.

Figure 3

Urn Problem 1.

The credibilities $q^{(\nu)}(I)$, for $I = 2$, 3 and 7, as functions of ν. $I = 3$ is the correct hypothesis. The a priori credibilities are: $q^{(0)}(2) = q^{(0)}(3) = 0.04545$, $q^{(0)}(7) = 0.54545$.



• *B. Second urn problem*

In the foregoing example, there was included in $\mathfrak{H}$ a hypothesis which is identical with the hidden law. In other words, there was an I for which $C^{(\nu)}(I)$ will become unity as $\nu \to \infty$, i.e., a perfect matching is realized for this I. In terms of A's of (4.15), this means that there is only one I for which $A(I)$ takes the maximum value and that this maximum value is equal to the empirical counterpart $\prod_{i=0}^{n} \alpha_i^{\alpha}$, see (4.15) and (5.4).

In the next example there is no hypothesis which corresponds exactly to the hidden law. That means there is no I for which $C^{(\nu)}(I)$ will become unity for $\nu \to \infty$. However, the experiment will be so arranged that there will be one I for which $A(I)$, of (4.9), will be larger than for any other $A(I)$'s. In this case, according to (4.18), the credibility of this particular I will become unity in spite of the fact that its confirmability never becomes unity.

The random number production is such that $I_0/10$ becomes 1/3, i.e., "0" and "1" are produced randomly at the ratio of 1 and 2. The set $\mathfrak{H}$ of hypotheses is the same as in the preceding example. We can reinterpret this new problem by assuming that the urn contains 30 balls, of which 10 are white. The hypotheses are artificially restricted to the following types:

$H(I)$: $3I$ balls out of the 30 are white and $30-3I$ balls are black. $I=0, 1, 2, \ldots, 10$.

Then, the hypothesis nearest to the truth will be $H(3)$, which means that there are 9 white balls. This $H(3)$ is expected to obtain credibility unity at large values of ν.

Under the circumstances described above, we have

$$\begin{aligned} &\alpha_0 = 1/3 \ , \quad \alpha_1 = 2/3 \ , \\ &p(0\,|\,3) = 0.3 \ , \quad p(1\,|\,3) = 0.7 \ , \\ &p(0\,|\,4) = 0.4 \ , \quad p(1\,|\,4) = 0.6 \ , \end{aligned} \tag{7.4}$$

from which follows

$$\begin{aligned} &A(3) = 0.52776 \ , \quad A(4) = 0.52415 \ , \\ &C^{(\infty)}(3) = 0.99593 \ , \quad C^{(\infty)}(4) = 0.98534 \ . \end{aligned} \tag{7.5}$$

Therefore, we can see that $H(3)$ and $H(4)$ are in a good competition. A slight fluctuation of $\alpha_0^{(\nu)}$ and $\alpha_1^{(\nu)}$, from the limiting value 1/3 and 2/3 can swing the balance between $q^{(\nu)}(3)$ and $q^{(\nu)}(4)$. In particular, if at a certain number ν, $C^{(\nu)}(3)$ and $C^{(\nu)}(4)$ happen to be equal, then $H(3)$ and $H(4)$ are equally well "confirmed." Therefore we can expect, as in formula (3.28), that for such a ν, the ratio of $q^{(\nu)}(3)$ to $q^{(\nu)}(4)$ simply becomes the ratio of the a priori credibilities, $q^{(0)}(3)$ and $q^{(0)}(4)$. In one of the experiments, in which $q^{(0)}(3)$ was put equal to 12/22 and all other a priori credibilities were put equal to 1/22, we happened to have at $\nu = 470$,

$$C^{(470)}(3) = 0.991504, \quad C^{(470)}(4) = 0.991535 \ . \tag{7.6}$$

Thus, $H(4)$ was slightly better "confirmed" than $H(3)$. At this point, the credibilities were

$$q^{(470)}(3) = 0.922382, \quad q^{(470)}(4) = 0.0776184 \ . \tag{7.7}$$

The ratio is 11.884, while $q^{(0)}(3)/q^{(0)}(4) = 12.000$. The slight difference in favor of $H(4)$ is due to the fact that $C^{(470)}(4)$ is larger than $C^{(470)}(3)$.

The experiments described in Fig. 6 is the case where all the a priori credibilities are equal, $q^{(0)}(I) = 1/11$. At $\nu = 1000$, the confirmabilities were found to be

$$C^{(1000)}(3) = 0.995035, \quad C^{(1000)}(4) = 0.986962 \ . \tag{7.8}$$

The discrepancy of these values from their respective theoretical limiting values is of the order of 0.1%. The credibilities of $H(3)$ and $H(4)$ at $\nu = 1000$ were

$$q^{(1000)}(3) = 0.995536, \quad q^{(1000)}(4) = 0.446419 \times 10^{-2} \ . \tag{7.9}$$

It goes without saying that the convergence $q^{(\nu)}(3) \to 1$ is much slower here than in the previous case. This can be seen by comparing Fig. 4 and Fig. 6.

• *C. Kochen's pattern-recognition problem*[11]

This problem is an attempt to make a computing machine guess a hidden pattern existing in a sequence of binary numbers. The items, i.e., the elements of $\mathfrak{D}$, are different binary numbers of five digits, $n = 32$. The number of possible patterns (hypotheses) are $2^n - 1 = 4{,}294{,}967{,}295$. But

Figure 4

Urn Problem 1.

The full lines represent the credibilities $q^{(\nu)}(I)$, for $\nu = 0$, 8, 32, 128, as functions of I. The weight shifts from $I=7$ to $I=3$, and finally $I=3$ only remains. The broken line gives the confirmability $C^{(\nu)}(I)$, at $\nu = 128$, (function of I). $C^{(\nu)}(I) = 0$ for $I=0$ and 10, for they are "logically excluded." At $\nu \to \infty$, both $q^{(\nu)}(3)$ and $C^{(\nu)}(3)$ become unity in this case.

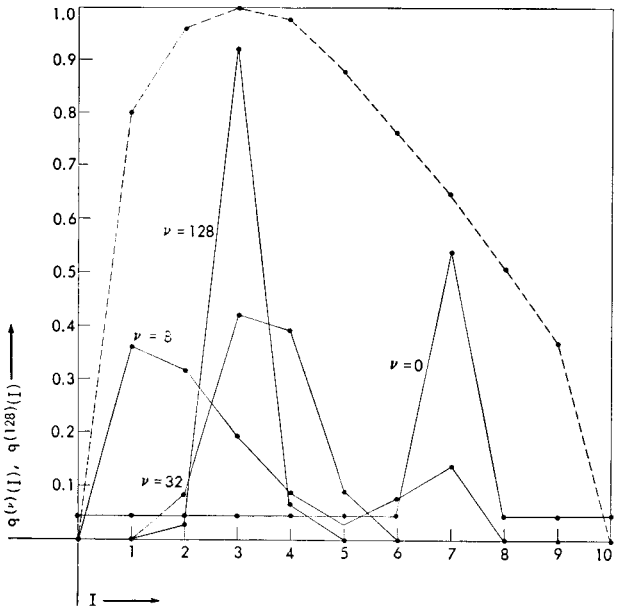


Figure 5

Urn Problem 1.

The inductive entropy function $U^{(\nu)}$ as a function of ν in an experiment, illustrating the "Inverse H-Theorem."

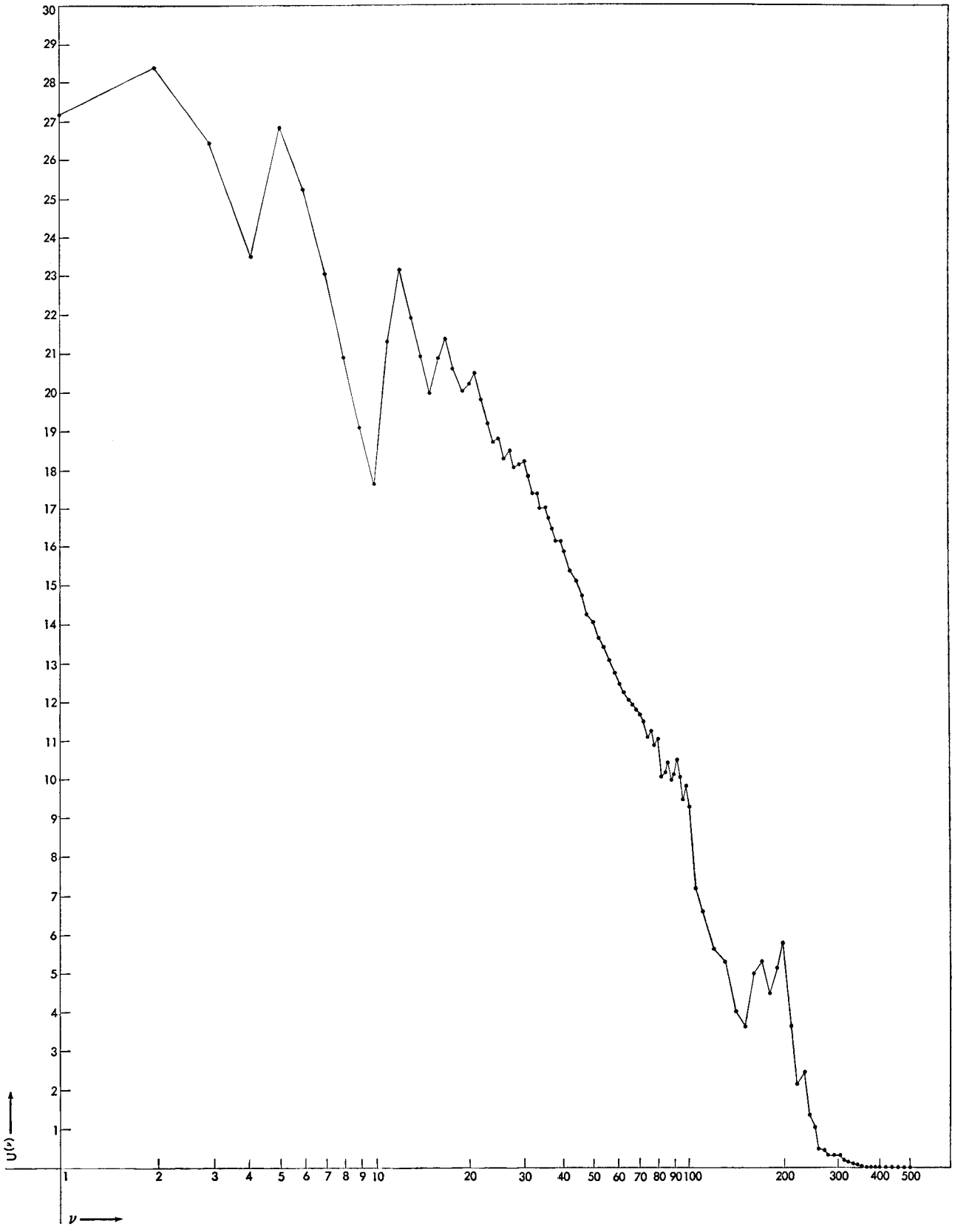


Figure 6

A case in which the credibility of one hypothesis tends to unity although its confirmability tends to a value less than unity.

(Unfortunately, this fact cannot be seen clearly on the scale used here.) The full lines are the credibilities, the broken line is the confirmability at $\nu = 1000$. Comparison with Fig. 4 will show how convergence is worse in this case.

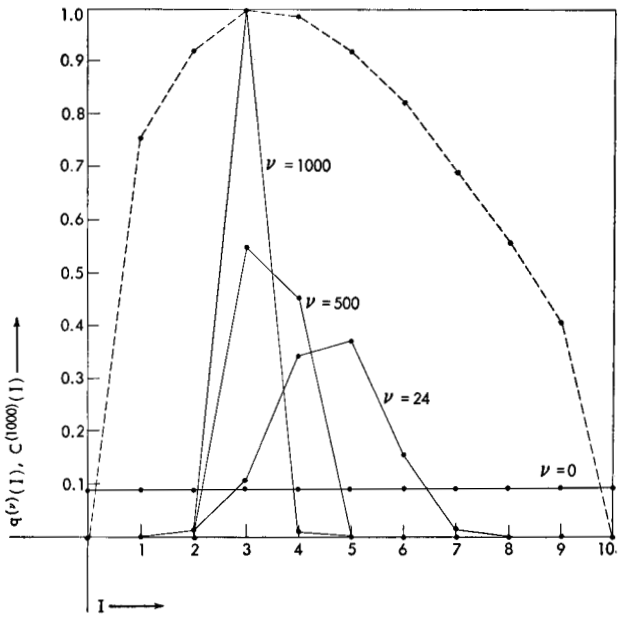
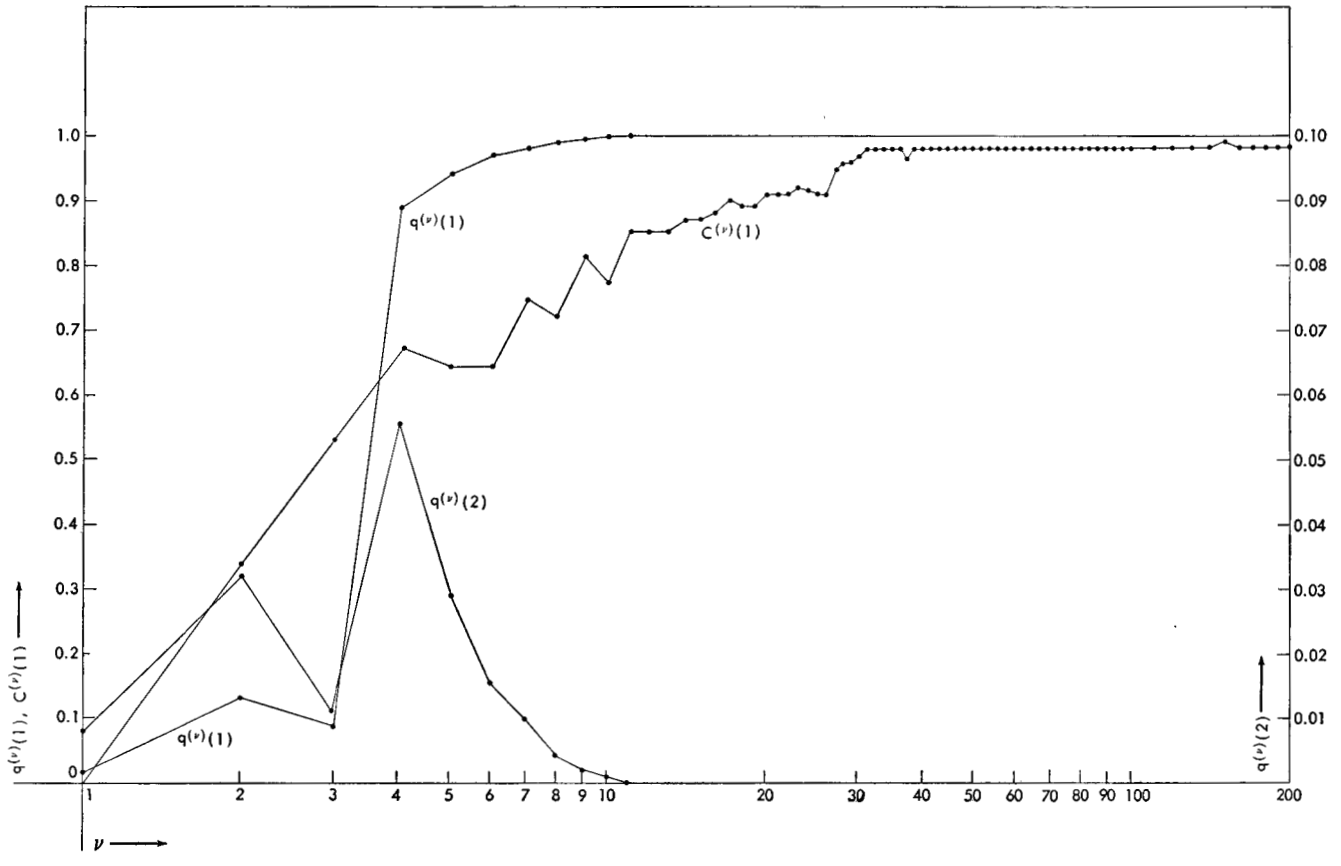


Figure 7

Kochen's pattern problem.

The correct hypothesis is $(X1XX0)$. The a priori credibility of each of 243 hypotheses is put equal to $1/243$. $q^{(\nu)}(1)$ and $C^{(\nu)}(1)$ are the credibility and the confirmability of the correct hypothesis $(X1XX0)$. $q^{(\nu)}(2)$ is the credibility of a wrong hypothesis $(XXXX0)$, which however cannot be "logically" excluded. The scale of $q^{(\nu)}(2)$ is amplified by factor 10 as compared with that of $q^{(\nu)}(1)$. The confirmability of the wrong hypothesis $(XXXX0)$ is exactly $3/4$ of $C^{(\nu)}(1)$, therefore is not entered in the chart.



Kochen's choice of patterns is such that a pattern can be expressed as a ternary number of five digits, in which a digit can be occupied by 0, 1 or X. The meaning of such a ternary number as a pattern is that the symbol X is allowed to take the value 0 or 1. For instance, (00X11) contains two items (00011) and (00111). This is very much the same as the idea used in the expression (2.17). The number of hypotheses becomes then $N=3^5=243$, which is still very large, but considerably smaller than 2^n-1. The dimension $W(I)$ of hypothesis I is 2^a, where a is the number of X's in the ternary expression of the hypothesis. Under the "homogeneity assumption," (2.3), the extent and intent of the hypothesis are given by $E(I)=\log W(I)=a$, and $\mathfrak{J}(I)=\log \frac{n}{W(I)}=(5-a)$.

Kochen's original problem is to make a computing machine guess the hidden hypothesis when the machine is shown various numbers, together with information as to whether the numbers do or do not belong to the hypothesis. Kochen's ingenious methods to make the machine behave intelligently in this guessing game are quite remarkable and are expected to be published before long. In Kochen's experiment, the machine is often shown a number which does not belong to the pattern and is informed to this effect. In our experiment, however, which is a model of the methods of natural sciences, the machine is shown only the numbers which do belong to the hypothesis. In order to conform with the homogeneity assumption, the numbers belonging to the hypothesis are shown to the machine in a random fashion with equal frequency for each number.

In the series of experiments described here, the true hypothesis (law) is $X1XX0$ which contains 8 different items, and these numbers are given to the machine at random with equal probabilities. The a priori credibilities $q^{(0)}(I)$, $I=1, 2, \ldots, 243$ are, in one experiment, assumed to be uniform, i.e., $q^{(0)}(I)=1/243$ for each I. In another experiment $q^{(0)}(I)=W(I)/\sum_{I'} W(I')$. But, as expected, this difference in the a priori probabilities is effaced very quickly as the observation accumulates. The credibility of the right hypothesis

Figure 8

Kochen's pattern problem.

The full lines represent the different paths by which the inductive entropy $U^{(\nu)}$ decreased in ten different runs. The a priori credibilities are $q^{(0)}(I)=1/243$. The broken line represents the average of these ten experimental values. The broken dotted line represents the average of ten other experimental values of $U^{(\nu)}$ based on the a priori probabilities $q^{(0)}(I)=W(I)/\sum_{I'} W(I')$. After $\nu=10$, there is practically no difference.

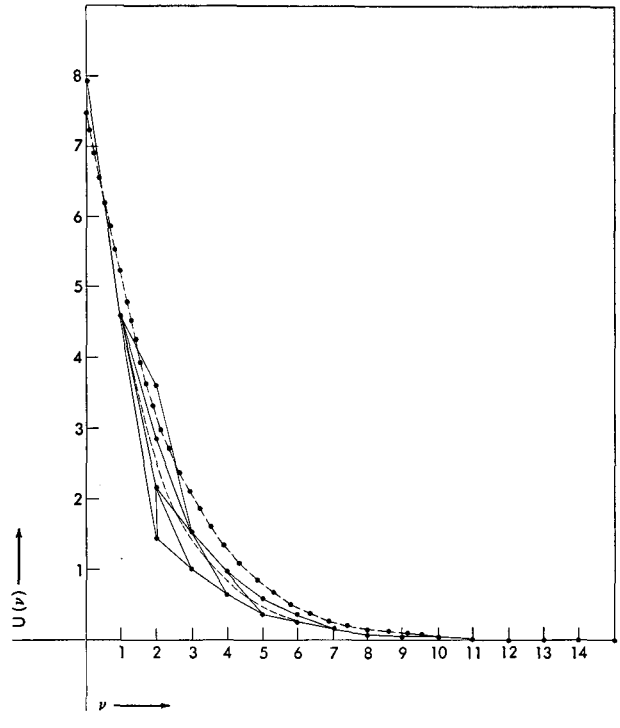
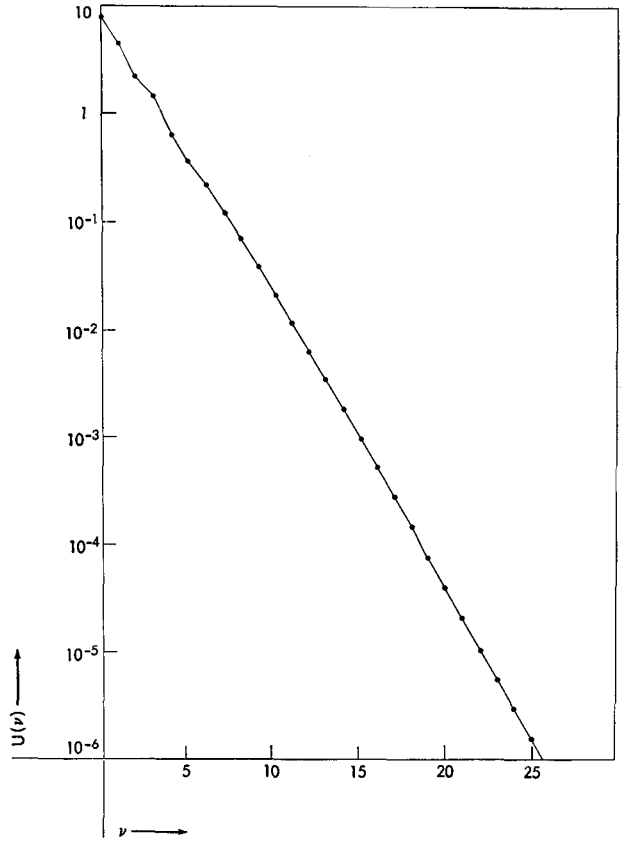


Figure 9

Kochen's pattern problem.

This graph shows the detail of the dependence of $U^{(\nu)}$ on ν, which looks almost like an exponential decrease. The a priori probabilities are $q^{(0)}(I)=1/243$. This is not an average curve; it represents the actual values of an individual run.



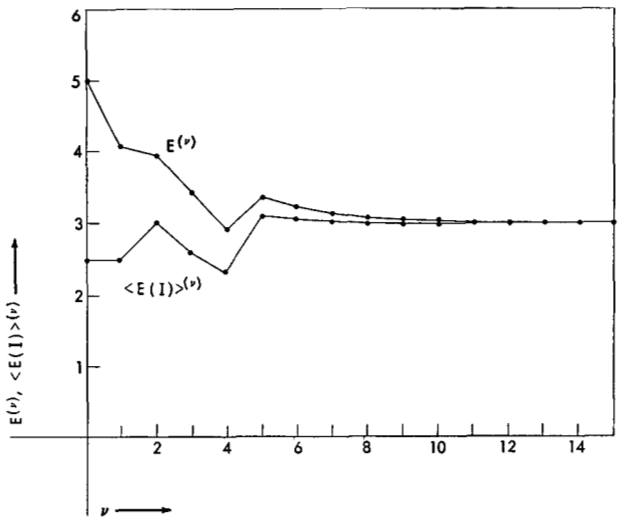


Figure 10

Kochen's pattern problem.

The quantities $E^{(\nu)}$, (6.21), and $\langle E(I)\rangle^{(\nu)}$, (6.23) plotted against ν. The a priori credibilities are $q^{(0)}(I) = W(I)/\sum_{I'} W(I')$. The theorem (6.25) is illustrated.

falls within 0.1% of unity after $\nu = 12$ in all the runs. The inductive entropy $U^{(\nu)}$ decreases almost exponentially after $\nu = 7$ and becomes less than 0.001 after $\nu = 15$ or 20. See (6.30). The confirmability is a stringent test for any hypothesis, and $C^{(\nu)}(I)$ of the correct hypothesis, i.e., $C(X1XX0)$ is still 0.983 at $\nu = 184$. But there is no doubt that $C^{(\nu)}(X1XX0)$ becomes unity at a very large ν. In the present experiment, hypothesis $(XXXX0)$ is not logically refuted, since all the specimens of $(X1XX0)$ are also specimens of $(XXXX0)$. This hypothesis $(XXXX0)$ belongs to $\mathcal{J}$ of the experimental data $\mathfrak{B}$ ensuing from the law $(X1XX0)$. In one experiment, $q^{(\nu)}$ of $(X1XX0)$ increased almost uniformly with ν, while $q^{(\nu)}$ of $(XXXX0)$ increased a little at lower values of ν but finally disappeared with further increasing values of ν. $C^{(\nu)}$ of $(XXXX0)$ is just 3/4 of $C^{(\nu)}$ of $(X1XX0)$, therefore remains finite. This means the credibility of $(XXXX0)$ becomes zero, while its confirmability remains finite. The product of confirmability and credibility may be a good conservative measure of the goodness of a hypothesis. The results are plotted in Fig. 7. The a priori probabilities of all the 243 hypotheses are set equal to 1/243. The hypothesis $I = 1$ is the correct one, i.e., $(X1XX0)$; $I = 2$ means a wrong hypothesis $(XXXX0)$.

The behavior of the inductive entropy is depicted in Figs. 8 and 9. No exception has been observed to the monotonous decrease of $U^{(\nu)}$ in any individual case in this problem.

The quantity $E^{(\nu)}$ defined in (6.32) and (6.33), and $\langle E(I)\rangle^{(\nu)}$ defined in (6.34) are plotted in Fig. 10. These quantities roughly correspond to the ignorance regarding the individual outcome at stage ν. The a priori credibilities here are given to be proportional to the dimension of each hypothesis, i.e., $q^{(0)}(I) = W(I)/\sum_{I'} W(I')$. This means, according to (6.33),

$$p^{(0)}(i) = \sum_{I} W(I)p(i|I) \Big/ \sum_{I'} W(I') \ , \tag{7.10}$$

where $p(i|I)$ is $1/W(I)$ if I includes i, and is zero otherwise. The numerator $\sum_{I} W(I)p(i|I)$ will then become a number of hypotheses including a given item, i.e., $2^5 = 32$. The denominator is equal to

$$\binom{5}{0}2^5 + \binom{5}{1}2^4 \cdot 2 + \binom{5}{2}2^3 \cdot 2^2 + \binom{5}{3}2^2 \cdot 2^3 + \binom{5}{4}2 \cdot 2^4 + \binom{5}{5}2^5 = (2+2)^5 = 1024 \ .$$

Hence, $p^{(0)}(i) = 1/32 = 1/n$. This satisfies the condition under which the theorem given in (6.36) was derived. We can indeed observe in Fig. 10 that relation (6.36) is satisfied. Both $E^{(\nu)}$ and $\langle E(I)\rangle^{(\nu)}$ converge, as was predicted, to the extent $E(I)$ of the correct hypothesis, which is 3.

Acknowledgment

The author would like to thank Manfred Kochen and Samuel Winograd for stimulating discussions he had with them concerning various problems closely related to the present work. It is also his pleasure to note that the critical comments by William Hanf were valuable in preparing the final version of the manuscript. Finally, he owes sincere thanks to William Kopka for carrying out beautifully the tedious job of programming and running on the IBM 704 the simulated experiments described in Section 7.

References and Footnotes

1. H. L. Gelernter, *Proceedings of the First International Conference on Information Processing, UNESCO*, Paris, 1959. (To be published).
2. A. L. Samuel, *IBM Journal of Research and Development*, **3,** 210 (1959).
3. It is well known that Carnap devoted a great deal of work to the subject of inductive probabilities—R. Carnap, *Logical Foundation of Probability*, Univ. of Chicago Press, 1950. The present paper does not necessarily conform with Carnap's theoretical framework.
4. N. Goodman, *Fact, Fiction and Forecast*, Harvard University Press, Cambridge, Mass., 1955.
5. S. Watanabe, *Reviews of Modern Physics*, **27,** 179 (1955). See, in particular, Section 6 on irreversibility of inference.
6. R. Ruyer has made some interesting remarks in this connection. See R. Ruyer, *Cybernétique et l'origine de l'information*, Flammarion, Paris, 1954.
7. Gaston Bachelard, "La psychanalyse du feu" (*Collection Psychologie, NRF*, Gallimard, Paris, 1938), p. 50.
8. For applications of this formula in physics, see S. Watanabe, *op. cit.*, p. 180.
9. The author's tnanks are due to Dr. Manfred Kochen for mentioning this interesting example to him.
10. S. Watanabe, *op. cit.;* in particular, Section 4.
11. Manfred Kochen, ms to be published.

Original manuscript received June 12, 1959
Revised manuscript received January 22, 1960