since it was proposed by Boole as a test of the "received" theory of probability as against his. (For the history see our *1986*, §6.2.) The problem is the following:

An event E can happen only as a result of either of two 'causes' A_1 or A_2. The probability of A_1 is c_1 and tht of A_2 is c_2. Also the probability of E, if A_1, is p_1 and that of E, if A_2, is p_2. Required the probability of E. In modern notation

$$\begin{aligned} \text{Given:}\quad & P(A_1) = c_1,\ P(A_2) = c_2 \\ & P(E \mid A_1) = p_1,\ P(E \mid A_2) = p_2, \\ & E \to A_1 \vee A_2, \\ \text{find:}\quad & P(E). \end{aligned}$$

MacColl's solution runs as follows. Since, by hypothesis, $E \leftrightarrow E(A_1 \vee A_2)$,

$$\begin{aligned} P(E) &= P(EA_1) + P(EA_2) - P(EA_1A_2) \\ &= P(A_1)P(E \mid A_1) + P(A_2)P(E \mid A_2) - P(A_1A_2)P(E \mid A_1A_2) \\ &= c_1p_1 + c_2p_2 - P(A_1A_2)P(E \mid A_1A_2). \end{aligned} \tag{1}$$

Boole's method yielded a solution depending only on the four parameters c_1, c_2, p_1, p_2, namely: the unique value of u in the probability range satisfying the equation

$$\frac{(u - c_1p_1)(u - c_2p_2)}{c_1p_1 + c_2p_2 - u} = \frac{(1 - c_1(1 - p_1) - u)(1 - c_2(1 - p_2) - u)}{1 - u} \tag{2}$$

MacColl attempts to prove Boole's solution to be wrong, but we have found his argument faulty (*Hailperin 1986*, 366–68). In a subsequent paper, *MacColl 1897*, he recurs to the problem and this time notes that the third term on the right in (1) "may have numberless values, all consistent with the data" whereas Boole's equation (2) produces only one value (in the probability range). He asserts: "The fallacy that vitiates Boole's whole reasoning on probability is to be found in his definition of independent events". This isn't entirely correct: the source of Boole's deviancy from "received" probability theory is his assumption that one can "ascend" to simple independent events representing the compound events of the problem, and by appropriately conditioning based on the logical relations of these compounds, reproduce the problem in terms of such conditioned events. When there are no logical relations, then the events are treated as independent (see §2.5 above).

Our solution to Boole's Challenge Problem is presented in §5.5 below.

Chapter 3

The Twentieth Century

§3.1. Confirmation—from Keynes to Carnap

Beginning with Francis Bacon (1561–1626), and continuing on up to the present, an array of writers concerned themselves with the question of inductive inference. The absence of a rational basis for such inferences was pointedly brought out by Hume (1711–1776). And, once it became clear that inductive inferences could not attain certainty, it was natural to think of introducing probabilistic notions. Perhaps the earliest attempts at linking probability with contingent inferences of this nature were those associated with Bayes' theorem and inverse probability. Here, as indicated in §1.6 above, the interest is in the probability of an event occurring *in a single trial* (Bayes) or on the *next* occasion (Laplace) if it has occured n times previously. This is to be contrasted with the more usual inductive inference which has a universal generalization for its conclusion. (In this connection it might be noted that the problem of how to associate probability with universally quantified propositions has, as yet, no recognized satisfactory solution.)

With regard to generalized inductive inferences we have the pessimistic opinion of C. D. Broad who in a paper on the 'logic of inductive inference' wrote (*1918*, 26):

> ... (1) that unless inductive conclusions be expressed in terms of probability all inductive inference involves a formal fallacy; (2) that the degree of belief which we actually attach to the conclusions of well-established inductions cannot be justified by any known principle of probability, unless some further premise about the physical world be assumed; and (3) that it is extremely difficult to state this premise so that it shall be at once plausible and non-tautologous.

Notwithstanding, there were, in the first half of the twentieth century, at least three efforts at a grand synthesis of foundations of probability and inductive inference. We shall only briefly discuss them—two in this section and the third in the next—confining our attention just to aspects relevant to our topic; we wish to distinguish our concept of probability logic from that of these efforts, which are often described as being a "logic of probable inference". To avoid possible misunderstanding, we emphasize that we are not judging the larger aims of these studies.

Keynes' *A Treatise on Probability (1921)* opens with a quotation from Leibniz which expresses the hope for a new kind of logic that would treat of degrees of probability. Apparently fulfilment of this hope was for Keynes a principal aim. But a great part of his *Treatise* is devoted to matters of philosophical analysis, to criticism of previous writings, to applications of probability, and to a justification of induction, little of this being related our study. Only slightly more than two of its 33 chapters have a connection with what we consider to be probability logic. Keynes characterizes this portion as follows (*1921*, 133):

> The object of this [chapter XII] and the chapters following is to show that all the usually assumed conclusions in the fundamental logic of inference and probability follow rigorously from a few axioms, in accordance with the fundamental conceptions expounded in Part I. This body of axioms and theorems corresponds, I think, to what logicians have termed the *Laws of Thought*, when they have meant by this something narrower than the whole system of formal truth. But it goes beyond what has been usual, in dealing at the same time with the laws of probable, as well as of necessary, inference.

What Keynes here refers to as "the fundamental conceptions expounded in Part I" may be briefly outlined as follows. It is asserted that all propositions are true or false.[1] Probability (less than certainty) only arises when there is an argument for a conclusion which does not follow demonstratively from the premises. If knowledge of h justifies a rational belief in a of degree α, it is said that there is a probability relation of degree α between a and h, and this is written '$a/h = \alpha$'.[2] In some cases this degree may be numerical, but need not be in general. Keynes does not consider

1. This assertion was disputed even in ancient times. See, for example, the summary of the early history of this doctrine in the Appendix 'On the history of the law of bivalence' of *Łukasiewicz 1930* (English translation in *Łukasiewicz 1970*, 176–78). In modern times quantum mechanics provides some grounds for believing that bivalence of propositions is not necessarily applicable to the physical world.

2. Admitting to its notational deficiencies Keynes nevertheless adopts it instead of '$P(a/h) = \alpha$' as being 'less cumbrous'.

the question of reconciling uniqueness of the degree of rational belief with the possibility of there being many *different* non-demonstrative arguments from one and the same given set of premises to a given conclusion—for example, the one-step inference 'h to a' is itself a non-demonstrative argument (except if the a happens to be a logical consequence of h). His (i.), p. 135, postulates uniqueness if h is not an inconsistent conjunction.

Many subsequent writers (though not all) refer to Keynes' concept as 'degree of confirmation' rather than as probability; or, as in Carnap (*1950*, 2nd edition *1962*) as a special kind of probability—'probability$_1$', in contrast to 'probability$_2$' which pertains to relative frequency. Sometimes it is referred to as 'epistemic' or as 'inductive' probability. Keynes' belief that the notion is fundamentally non-numerical has had little following and many writers drop this feature; similarly for his epistemic-mentalistic description in terms of knowledge and belief. Nicod (*1930*, 207), for example, simply refers to a Keynes relation of probability as a *probable inference*. We discuss his few brief remarks about it.

The material occurs at the beginning of Nicod's lengthy essay on induction, which is taken as a species of inference. Two kinds of inference are distinguished (*Nicod 1930*, 207):

> We first regard inference as the perception of a connection between the premises and conclusion which asserts that the conclusion is true if the premises are true. This connection is implication, and we shall say that an inference grounded on it is a *certain inference*. But there are weaker connections which are also the basis of inferences. They have not until recently received any universal name. Let us call them with Mr. Keynes *relations of probability* (*A Treatise on Probability*, London, 1921, ch. 1.). The presence of one of these relations among the group A of propositions and the proposition B indicates that in the absence of any other information, if A is true, B is probable to a degree *p*. A is still a group of premises, B is still a conclusion, and the perception of such a relation between A and B is still an inference: let us call this second kind of inference *probable inference*.

Nicod's phrase "in the absence of any other information" we take to mean that only the premises in A (or A, if a single premise) are to be considered. This is the normal way one thinks of inference. His understanding of a Keynes probability relation, $P(B/A) = p$, is that 'if A is true, B is probable to a degree p'. We take 'indicates' to mean 'implies'. In his discussion immediately following, the 'A is true' becomes 'A is certain'. Nicod is thus affirming the correctness of the inference form

$$\frac{P(B/A) = p,\ A \text{ is certain}}{B \text{ is probable to a degree } p}$$

or, if we write it to accord with Keynes' view that probability is a two-argument notion,

$$\frac{P(B/A)=p,\ P(A/t)=1}{P(B/t)=p}$$

where 't' stands for a logically true sentence. But affirming the correctness of this inference form and nothing else, tells us little about $P(B/A)=p$. (Compare, for example, a proposed explanation of the ordinary conditional $C(A,B)$ in propositional logic which says that it indicates

$$\frac{C(A,B),\ A}{B}.$$

Here either AB, $A\to B$, or $A\leftrightarrow B$ are all interpretations for $C(A,B)$ which produce a valid inference form.)

Nicod then goes on to the case of probable premises (*1930*, 208):

> We have so far considered only premises that are certain. But any inference which yields something, starting from premises taken only as certain, still yields something starting from premises taken only as probable, and this holds for both types of inferences, certain and probable. We can even assert that starting from premises which, taken together have a probability p, a certain inference will confer on its conclusion the same probability p; and a probable *inference*, which would confer on its conclusion the probability q if its premises were certain, will confer on its conclusion the probability pq.

Both of the assertions in this quotation are incorrect. In the first place a certain (i.e., necessary) inference having a premise with probability p,[3] confers on its conclusion not probability p but probability *not less that* p. (For a partial history of this error and its correction by ~~Bolzano,~~ see *Hailperin 1988b*, §§4–5, also §2.1 above). As for the second assertion, the claim is

$$\frac{P(B/A)=q,\ P(A/t)=p}{P(B/t)=pq}.$$

The conclusion here can't be valid since by the multiplication rule (*Keynes 1921*, 135),

$$pq=P(B/At)(P(A/t)=P(AB/t).$$

However, since $P(B/t)\geq P(AB/t)$, it would be valid if the conclusion were

$$P(B/t)\geq pq.$$

3. This could mean that the premises of the inference are either '$P(B/A)=1$, $P(A)=p$', or '$P(A\to B)=1$, $P(A)=p$'. The difference is immaterial since, for $p>0$, '$P(B/A)=1$' is equivalent to '$P(A\to B)=1$'. (See Theorem 5.36 below.)

We return to Keynes' *Treatise*. In chapter 12 he presents an axiomatic formulation for his a/h notion. But what one finds there could not pass muster by current standards. For example, there is no listing of undefined notions, no specification of proper syntactic structure, and no explicit rules of inference; further, some definitions are conditional (so that eliminability is in doubt), while some lack axioms to make them operative. Rather than introducing this system—which would require extensive emendatory discussion—we shall instead present what we believe to be an adequate surrogate, namely the axiomatization of 'confirmation' due to J. Hosiasson(-Lindenbaum) (*1940*, 133):

> Let a, b, and c be variable names of sentences belonging to a certain class,[1] the operations $a.b$, $a+b$, and $\overline{a}$, the (syntactical) product, the sum and the negation of them. Let us further assume the existence of a real non-negative function $\mathfrak{c}(a,b)$ of a and b, when b is not self-contradictory. Let us read '$\mathfrak{c}(a,b)$' 'degree of confirmation of a with respect to b' and take the following axioms:
>
> Axiom I. If a is a consequence of b, $\mathfrak{c}(a,b)=1$.
>
> Axiom II. If $\overline{a.b}$ is a consequence of c,
>
> $$\mathfrak{c}(a+b,c)=\mathfrak{c}(a,c)+\mathfrak{c}(b,c).$$
>
> Axiom III. $\mathfrak{c}(a.b,c)=\mathfrak{c}(a,c)\cdot\mathfrak{c}(b,a.c)$.
>
> Axiom IV. If b is equivalent to c, $\mathfrak{c}(a,b)=\mathfrak{c}(a,c)$.[2]
>
> As may easily be seen, the interval of variation of $\mathfrak{c}$ is $(0,1)$; this is quite conventional.
>
> [1]The class must be broad enough to include all sentences for which we desire to speak about confirmation.
>
> [2]These axioms are analogous to St. Mazurkiewicz's system of axioms for probabilities (see *Zur Axiomatik der Wahrscheinlichkeitsrechnung, Comptes rendus des séances de la Société des Sciences et des lettres de Varsovie*, vol 25 (1932)).

What is here referred to as 'a certain class' is not narrowly specified, but by virtue of the presence of logical sum, product, and negation in the axioms it is clear that the class would have to be closed under these operations. Conceivably the sentences of the class could involve additional logical structure (e.g., quantifiers) but this can play no role in confirmation theory (as based on these axioms) since the axioms show no interaction of the $\mathfrak{c}$ function except with propositional connectives. Although, by Axiom IV, provision is made for replacement in the second argument place of the $\mathfrak{c}$ function of one (logically) equivalent sentence by another, there is none justifying replacement in the first position. We take this to be an oversight and shall assume that Axiom IV is modified so as to include this provision. Subsequent to this axiomatization by Hosiasson other formulations of confirmation have appeared. However, the differences between these and Hosiasson's are minor and have no relevance to our study.

Let t be some fixed logically true sentence. Then $\mathfrak{c}(a, t)$ is a one-place function of the sentence variable 'a'; if $\mathfrak{c}(a, t) \neq 0$ one readily proves from axioms III, I, and IV, that

$$\mathfrak{c}(b, a) = \mathfrak{c}(a.b, t)/\mathfrak{c}(a, t). \tag{1}$$

Thus the two-place $\mathfrak{c}(b, a)$ is expressible in terms of a one-place function. It is readily seen that when the formal $\mathfrak{c}(a, t)$ is interpreted as a probability, $P(a)$, and (1) taken to be a definition of conditional probability, $P(b \mid a)$, the above axioms for confirmation become correct assertions about conditional probability. Thus the axioms of Hosiasson are insufficient to characterize anything different from conditional probability. If, as in the opinion of some writers, the confirmation notion is central to inductive logic, something more needs to be added.

During the 1940s two separate approaches were initiated to define confirmation as a syntactic notion so as to provide a basis for inductive logic. For our purposes it will suffice to sketch one of these—that due to Carnap—as it is the one which was extensively developed and is widely known. (For the other, see *Hempel 1943*.)

Unlike Keynes and Hosiasson who, in connection with confirmation, recognize that sentences are combinable by sentential connectives but say nothing further about the syntax of the language, Carnap assumes that he has a specific (type of) language for which his explicatum for the informal probability$_1$, generically denoted '$\mathfrak{c}(h, e)$', is to be defined. He contends that the notion should be 'L-determinate', i.e., depend only on the logical structure of h and e. Initially only simple languages are considered. The ultimate goal is a definition of the notion in a language adequate for science. In what follows we shall only be referring to his $\mathcal{L}_N$, a predicate language over a universe of N individuals, hence effectively without quantifiers. The language $\mathcal{L}_N$ has the N individual symbols $a_1, \ldots, a_N$.

Carnap acknowledges that an explicatum for probability$_1$ should have the commonly accepted general properties of confirmation (as given, for example, in Hosiasson's axioms). A function having these properties he calls a regular $\mathfrak{c}$-function. However, he argues that a much narrower notion is needed for inductive logic (*Carnap 1950*, or *1962*, 344):

> If the axioms of a system hold for all regular $\mathfrak{c}$-functions, then that system represents only a very small part of the theory of probability$_1$. This part, it is true, is of great importance because it contains the fundamental relations between $\mathfrak{c}$-values. But its weakness becomes apparent from the following facts. Let e and h be factual sentences in $\mathcal{L}_N$ such that e L-implies neither h nor $\sim h$. Then a theory of the kind mentioned does not determine the value of $\mathfrak{c}(h, e)$. Moreover, it does not even impose any restricting conditions upon this value: the assignment of any arbitrarily chosen real number between 0 and 1 is compatible with the theory. ... A theory of this kind states merely relations between $\mathfrak{c}$-values; thus if some $\mathfrak{c}$-values are given, others can be computed with the help of the theorems. There is an analogous restriction in the theory of probability$_2$; however, here the restriction is necessary. The statement of a particular value of probability$_2$ for two given properties is, in general, a factual statement (10B). Therefore, a logicomathematical theory of probability$_2$ cannot yield statements of this kind but must restrict itself to stating relations between probability$_2$ values. On the other hand, in the case of a theory of probability$_1$, there is no reason for this restriction. A sentence of the form '$\mathfrak{c}(h, e) = r$' is not factual but L-determinate. Therefore, a logicomathematical theory of probability$_1$, in other words, a system of inductive logic, can state sentences of this form. The fact that the axiom systems for probability$_2$ restrict themselves to statements which hold for all regular $\mathfrak{c}$-functions make these systems unnecessarily weak.

At the time he wrote this Carnap thought that an adequate explicatum for probability$_1$ would be something like his function $\mathfrak{c}^*$, described in the appendix to his *1950*. He later acknowledged that this was not satisfactory, that $\mathfrak{c}^*$ was only one in a whole continuum of possible $\mathfrak{c}$-functions and then, later, that even these were not entirely satisfactory as a basis for inductive logic (*Carnap and Jeffreys 1971*, 1). Nevertheless, by describing his $\mathfrak{c}^*$ (for $\mathcal{L}_N$) we can illustrate what sort of conception Carnap had in mind for his inductive logic, and to compare it with our notion of probability logic.

For Carnap a $\mathfrak{c}$-function is defined in terms of a measure m over sentences of the language, namely by setting

$$\mathfrak{c}(h, e) = m(e.h)/m(e) \quad \text{for } m(e) \neq 0.$$

For simple languages $\mathcal{L}_N$ specifying the values m for state-descriptions (basic conjunctions of atoms and negated atoms) is sufficient, these values being positive real numbers whose sum taken over all the state-descriptions equals 1. All other values are then determined via additivity of the measure. To obtain his $\mathfrak{c}^*$ Carnap specifies a particular measure, m^*, which is (i) 'symmetric' and (ii) has the same value for all 'structure-descriptions' of $\mathcal{L}_N$. A symmetric m function is one which assigns equal values to isomorphic state-descriptions—two state-descriptions being isomorphic if one is obtainable from the other by a permutation of the individual symbols, plus rearranging of conjunctions. A structure-description (determined by a state-description) is the disjunction of all state-descriptions isomorphic to it, arranged in lexicographical order. The value of m^* for a state-description S_i is then given by

$$m^*(S_i) = 1/\tau\zeta_i$$

when τ is the number of state-descriptions of $\mathcal{L}_N$ and ζ_i the number of state-descriptions isomorphic to S_i.

For our purposes we need not entertain this much complexity—it will suffice to limit ourselves to an $\mathcal{L}_1$, a language with one individual, in which case permuting individual symbols is irrelevant and a structure-description is the same as a state-description. We take this $\mathcal{L}_1$ as having the indivdual symbol 'a' and three one-place predicate symbols 'Q_1', 'Q_2', and 'Q_3'. Let $A_i = Q_i(a)$, $(i = 1, 2, 3)$. Then there are only 2^8 logically inequivalent sentences constructible for this language, namely those obtained by taking a disjunction of some, none, or all of the eight state-descriptions

$$A_1A_2A_3,\ \overline{A}_1A_2A_3,\ A_1\overline{A}_2A_3,\ \ldots,\ \overline{A}_1\overline{A}_2\overline{A}_3. \tag{2}$$

To each of these state-descriptions Carnap's theory would assign an equal measure 1/8. The one-argument $\mathfrak{c}^*(h,t)$,which by definition is $m^*(h)$, would equal $n(h)/8$, where $n(h)$ is the number of these which imply h. This value is independent of any interpretation of a, Q_1, Q_2, Q_3 and accords with Carnap's conception of (inductive) probability as a measure of closeness to validity based on the syntactic structure of sentence in a given language.

In contrast, probability logic as we are conceiving it is semantically rather than syntactically based. Generality and independence of interpretation are obtained by allowing any assignment of real numbers $k_1, \ldots, k_8$ to the state-descriptions (2), subject only to the requirement that they be non-negative and sum to 1. Specific application determines what particular set of real numbers to choose. On what basis one chooses these eight values, be it statistical, subjective, or merely hypothetical, is not a relevant matter for (pure) probability logic, any more than the basis on which one chooses to assign truth values to sentences when applying verity logic. Carnap would refer to our notion as being 'factual'. We think 'semantic' would be more appropriate as we believe '$P(h \mid e) = p$' to be analogous to the semantic sentence "The truth value of 'Snow is white' is truth" rather than to the factual sentence 'Snow is white'. Our ideas of a probability logic are more fully described in chapter 4 below.

We turn now to a distinctively different approach which also aims at justifying induction via a theory of probability and which its originator also referred to as a 'probability logic'.

§3.2. Reichenbach: probability as multivalued logic

Beginning with his dissertation of 1915, and continuing on through to the 1950s, H. Reichenbach wrote extensively on philosophical and technical matters connected with foundations of probability. Our references will be confined to his *The Theory of Probability (1949)*, a compendium of his settled views.

Reichenbach's notion of probability is, like Keynes', relational though conceptually different. Whereas Keynes thinks of it as a two-argument relation having a (possibly but not necessarily numerical) degree, Reichenbach takes it to be a three-argument relation with two of these arguments being classes of a special kind and the third a real number. For his 'fundamental probability relation' he writes

$$(i)(x_i \in A \underset{p}{\supset} y_i \in B), \tag{1}$$

where '$\underset{p}{\supset}$' is referred to as 'probability implication'. The classes A and B are *thought of* as having members x_1, x_2, ..., x_i, ... and y_1, y_2, ..., y_i, ... respectively. These are assumed to be denumerably ordered, and (1) is taken to mean that a probability implication relation holds for all corresponding pairs (x_i, y_i). It is also read as 'if an $x_i \in A$, then $y_i \in B$ with probability p'. Since the universal quantifier '(i)' and the inner structures of A and B play no role in his listing of the formal properties of '$\underset{p}{\supset}$' Reichenbach simplifies (1) to

$$A \underset{p}{\supset} B$$

or, later on, it is written

$$P(A, B) = p.$$

This latter expression is presented as a whole without any meaning attached to the separate parts, i.e., to '=', or '(A, B)', or '$P(A, B)$'. To avoid the typographical inconvenience of having complex numerical expressions such as '$\phi(q, r)$' underneath the (probability) implication symbol Reichenbach uses the form

$$(\exists p)(A \underset{p}{\supset} B)(p = \phi(q, r)).$$

(This is how Reichenbach writes it; for a correct rendering the scope of the quantifier should be the entire conjunction, not just the first member.)

We shall first state Reichenbach's axioms and then briefly describe their content (*1949*, 54–62):

Axiom I. $\quad p \neq q \supset [(A \underset{p}{\supset} B).(A \underset{q}{\supset} B) \equiv (\overline{A})]$

Axiom II1. $(A \supset B) \supset (\exists p)(A \underset{p}{\supset} B).(p = 1)$

Axiom II2. $\overline{(\overline{A})}.(A \underset{p}{\supset} B) \supset p \geq 0.$

Axiom III. $(A \underset{p}{\supset} B).(A \underset{q}{\supset} C).(A.B \supset \overline{C}) \supset (\exists r)(A \underset{r}{\supset} B \vee C).(r = p+q)$

Axiom IV. $(A \underset{p}{\supset} B).(A.B \underset{u}{\supset} C) \supset (\exists w)(A \underset{w}{\supset} B.C).(w = p \cdot u)$

According to Axiom I the same probability implication from A to B holds for two distinct reals if and only if A is empty. (By his conventions '$(\overline{A})$' means '$(i)(x_i \in \overline{A})$', i.e., that A has no members). By Axiom II1, the p for $A \underset{p}{\supset} B$ is 1 if A is contained in (implies) B; and by II2, $p \geq 0$ if it is not the case that A is empty. Axioms III and IV are the usual addition and multiplication rules, appropriately framed for the context.

In chapter 4 of his book Reichenbach 'constructs' a frequency interpretation for his $A \underset{p}{\supset} B$ in terms of postulated properties of sequences associated with A and B. However one can give a more general and simpler interpretation: Let P be a probability function defined on a Boolean algebra which is strictly positive (as well as normed to 1 and additive). One readly verifies that Reichenbach's axioms I–IV are satisfied if A and B are arbitrary elements of the Boolean algebra and '$A \underset{p}{\supset} B$' means

$$P(A.B) = p \cdot P(A),$$

i.e., that the conditional probability of B given A is p, but in this variant form rather than the more usual quotient form. (In Axiom I, since in our interpretation P is assumed to be strictly positive, the clause '$(\overline{A})$', i.e., A is empty, is equivalent to '$P(A) = 0$'.) Thus Reichenbach's axioms for his probability implication, as Hosiasson's for confirmation, need supplementation to characterize a notion which is distinguishable from conditional probability based on a one-argument probability function.

Of particular interest for our study is Reichenbach's claim (*1949*, chapter 10) to have constructed a probability logic with continuously many truth values. We believe this claim to be unsubstantiated. Our reasons are listed below after the following discussion.

Reichenbach introduces the notation 'hz_i' for '$z_i \in A$' and 'fx_i' for '$x_i \in B$' and converts his probability form

$$(i)(z_i \in A \underset{p}{\supset} x_i \in B)$$

to

$$(i)(hz_i \underset{p}{\supset} fx_i), \tag{2}$$

which is then written

$$P(hz_i, fx_i) = p. \tag{3}$$

But if read literally this latter form, containing free 'i', would be saying that P maps the i-th pair of corresponding members of the two sequences into p, which is clearly not intended since Reichenbach's interpretation for $P(hx_i, fy_i)$ involves the entire sequences (see (8) below.) Hence a more appropriate notation would be

$$P(\hat{h}, \hat{f}) = p, \tag{3'}$$

where $\hat{h} = \{\, hx_i : i \in \omega \,\}$ and $\hat{f} = \{\, fy_i : i \in \omega \,\}$ are the entire sequences; this would accord with his earlier notation $P(A, B) = p$.

Viewing (3) [better would be (3′)] as an assertion about the two propositional sequences Reichenbach says (p. 395): "The frequency interpretation [of (2) or (3)] is contructed by counting the number of true propositions 'fx_i' within the subsequence selected by the true propositions 'hz_i'." When hz_i is true for each i, Reichenbach writes

$$P(fx_i) = p \tag{4}$$

with frequency interpretation

$$P(fx_i) = \lim_{m \to \infty} \frac{1}{m} \overset{m}{\underset{i=1}{\mathrm{N}}} \{V[fx_i] = 1\}, \tag{5}$$

whose meaning is that $P(fx_i)$ is the limit of the relative frequency of the number of true occurrences of fx_i in the sequence. (Note that the right-hand side of (5) has no free 'i'.)

Since the probability logic Reichenbach intends to present deals with propositional sequences rather than propositions, he wishes to define operations for propositional sequences analogous to those for propositions. All goes well for those that correspond to the usual connectives. Using parentheses about the general term of a sequence to denote the sequence (hence binding its free 'i') he defines (p. 398):

$$\begin{aligned} (fx_i) \vee (gy_i) &=_{Df} (fx_i \vee gy_i) \\ (fx_i).(gy_i) &=_{Df} (fx_i.gy_i) \\ \overline{(fx_i)} &=_{Df} (\overline{fx_i}). \end{aligned} \tag{6}$$

These are legitimate definitions of operations on propositional sequences which result in propositional sequences and their P values would be given by (5). (These values ought to be written '$P((fx_i))$'.) But then he goes on with (p. 399, display numberings changed so as to mesh with ours):

We must now introduce a new propositional operation that will allow us to write degrees of relative probabilities. Since probabilities of this kind are written in the form $P(fx_i, gy_i)$, the content of the parentheses in this expression can be regarded as a compound proposition, resulting from the two components by a propositional operation, denoted by the comma. We shall call it *operation of selection*, or *comma operation*. We can also put the comma between propositional sequences. We then define, by analogy with those of (6),

$$(fx_i), (gy_i) =_{Df} (fx_i, gy_i). \tag{7}$$

The frequency interpretation of $P(fx_i, gy_i)$ is given by the expression

$$P(fx_i, gy_i) = \lim_{n\to\infty} \frac{\overset{n}{\underset{i=1}{\mathrm{N}}}\{V[fx_i.gy_i] = 1\}}{\overset{n}{\underset{i=1}{\mathrm{N}}}\{V[fx_i] = 1\}}. \tag{8}$$

This form makes clear why we speak of the operation of selection. The proposition fx_i selects the subsequence in which we count the frequency of gy_i. Since (7) allows us to regard the comma on the left side of (8) as standing between the propositional sequences, we can also say that the comma operation represents a selection from one sequence by another sequence.

The argument here lacks cogency. Reichenbach says that the content of the parentheses in '$P(fx_i, gy_i)$' can be regarded as a compound proposition. This seems to indicate that he is thinking of the 'fx_i' and 'gy_i' appearing therein as propositions, i.e., as instances of the propositional sequences. But, as we have earlier noted and as (8) shows, they have to be propositional sequences, not propositions. Furthermore, he seems to think that the selection idea provides him with a sequence whose general term is 'fx_i, gy_i'. But the selection only provides a term for *some* values of i, namely those for which fx_i is true, and nothing when fx_i is false. Hence no meaning for 'fx_i, gy_i' for all i if fx_i has false instances.

In his truth tables Reichenbach uses probability values in place of the two truth values of ordinary two-valued logic. These are probabilities of propositional sequences. However, unlike for the two-valued logic, where the value of a compound depends on the values of the two components, say fx_i and gy_i, he uses the three values

$$P(fx_i), \quad P(gy_i), \quad P(fx_i, gy_i),$$

to specify the value of a compound of fx_i and gy_i. (We are following Reichenbach's convention of omitting the parentheses around a general term of a propositional sequence allowing the general term to stand for the sequence.) This is how his truth tables appear (p. 400), abbreviated by the omission of the columns for '$\supset$' and '$\equiv$'):

Table1. Truth tables of probability logic

A. Negation

$P(f)$	$P(\overline{f})$
p	$1-p$

Restrictive Conditions:

1. $\frac{p+q-1}{p} \leq u \leq \frac{q}{p}$
2. $P(f, f) = 1$

B. Binary Operations

$P(f)$	$P(g)$	$P(f, g)$	$P(f \vee g)$	$P(f.g)$	$P(g, f)$
p	q	u	$p+q-pu$	pu	pu/q

Reichenbach considers these tables to be a sufficient basis for a probability logic. But let us compare them with the analogous ones for two-valued logic. We use 'V' for 'the truth value of'.

$V(\phi)$	$V(\psi)$	$V(\overline{\phi})$	$V(\phi \vee \psi)$	$V(\phi.\psi)$
p	q	$1-p$	$\max\{p, q\}$	pq

Here p and q are variables ranging over the set $\{0, 1\}$, with 0 and 1 having their usual arithmetic properties. Another way of expressing this is by the equations

$$\begin{aligned} V(\overline{\phi}) &= 1 - V(\phi) \\ V(\phi \vee \psi) &= \max\{V(\phi), V(\psi)\} \\ V(\phi.\psi) &= V(\phi) \cdot V(\psi). \end{aligned} \tag{9}$$

Since a (two-valued) logical expression is constructed out of finitely many basic elements (atomic propositions) its V value is determined in terms of those of the basic elements by a recursive use of (9). This enables two-valued logic to have a (an effective) notion of logical consequence. That a formula ψ is a logical consequence of a set of premises $\phi_1, \ldots, \phi_n$ can be determined by examining all possible assignments of 0 or 1 to the basic elements common to $\phi_1, \ldots, \phi_n, \psi$. But Reichenbach's tables provide no means of reducing P values of expressions to P values of simpler expressions nor, for that matter, is there a clear specification of the syntax of the language. We do not even know if, for example, '$\phi \vee (\psi, \chi)$' or '$(\phi, \psi).\chi$' are meaningful.

Since Reichenbach has

— no adequate justification for including 'fx_i, gy_i' in his language
— no complete specification of the syntax of his language
— no provision for general determination of the probability value of an expression in terms of its components, and
— no definition of logical consequence

we do not, on the basis of our understanding of the concept, accept his claim to have constructed a probability logic.

§3.3. Probabilistic inference revived

Despite De Morgan's cogently argued advocacy for probable inference as a part of formal logic, and Boole's conceiving of, and if not succeeding at least showing the feasibility of, a general method for obtaining such inferences, little of substance ensued. As evidenced by the writers we have mentioned in our preceding two sections, the interest was in justifying methodological principles of science. Not until we come to Suppes' *1966* is there something for our historical account. And even in his paper the topic of probabilistic inference was not central; it arose in the context of a criticism of employing the concept of 'total evidence' as a means of resolving the 'statistical paradox'. The paradox is stated in Suppes' opening paragraph (*1966*, 49):

> My purpose is to examine a cluster of issues centering around the so-called statistical syllogism and the concept of total evidence. The kind of paradox that is alleged to arise from uninhibited use of the statistical syllogism is of the following sort.
>
> (1) The probability that Jones will live at least fifteen years given that he is now between fifty and sixty years of age is r. Jones is now between fifty and sixty years of age. Therefore, the probability that Jones will live at least fifteen years is r.
>
> (2) The probability that Jones will live at least fifteen years given that he is now between fifty-five and sixty-five years of age is s. Jones is now between fifty-five and sixty-five years of age. Therefore, the probability that Jones will live at least fifteen years is s.
>
> The paradox arises from the additional reasonable assertion that $r \neq s$, or more particularly that $r > s$.

If we write the inference forms involved in Suppes' statement as

$$\frac{\begin{array}{c} P(A\,|\,B) = r \\ B \end{array}}{P(A) = r} \quad (1) \qquad\qquad \frac{\begin{array}{c} P(A\,|\,C) = s \\ C \end{array}}{P(A) = s} \quad (2)$$

then the rule of total evidence considers it illegitimate to use B and C separately as evidence when it is know that $B \wedge C$ is the case.

Aside from the difficulties associated with the rule of total evidence (see the criticism by Suppes in his paper on p. 50) there is the question as to what an inference form such as (1) means. As we have seen in our §3.1, Nicod considered (1)—with the major premise taken as a statement of a Keynes probability relation—to be a way of explaining Keynes' notion, thus implying that (1) was for him clearly valid. On the other hand Suppes says (p. 58):

> ... Now let us schematize this inference [i.e., (1)] in terms of *hypothesis* and *evidence* as these notions occur in Bayes' theorem
>
> $$\frac{\begin{array}{l} P(\text{hypothesis}\,|\,\text{evidence}) = r \\ \text{evidence} \end{array}}{\therefore P(\text{hypothesis}) = r},$$
>
> and the incorrect character of this inference is clear. From the standpoint of Bayes' theorem it asserts that once we know the evidence, the posterior probability $P(H\,|\,E)$ is equal to the prior probability $P(H)$, and this is patently false.

We have then two diametrically opposed opinions on the validity of (1). This inference form—with the $P(A\,|\,B)$ taken in the general sense of confirmation—was the subject of a lengthy essay by H. Kyburg. The essay was followed by a discussion involving seven participants of the colloquium at which it was presented (*Lakatos 1968*, 98–165). The positions on the inference were so widely diverse that we may reasonably conclude that it has no clear meaning.

Replacing the second premise in (1) by the probability statement '$P(B) = 1$' does produce the valid form

$$\frac{\begin{array}{r} P(A\,|\,B) = r \\ P(B) = 1 \end{array}}{\therefore P(A) = r} \quad (3)$$

But (3) is not usable for inferences of the kind in the statistical paradox, for '$P(B) = 1$' is true if and only if B is the sure event, and false otherwise.

We turn to some probabilistic inferences which Suppes presents. These are all clearly valid forms, since they are consequences of basic probability principles.

In place of the disputed (1) Suppes considers the natural generalization of the rule of detachment for probable inference to be

$$\frac{\begin{array}{c} P(A \mid B) = r \\ P(B) = \rho \end{array}}{\therefore P(A) \geq r\rho} \tag{4}$$

or, more generally,

$$\frac{\begin{array}{c} P(A \mid B) \geq r \\ P(B) \geq \rho \end{array}}{\therefore P(A) \geq r\rho} \tag{5}$$

He contrasts this with the one which uses the ordinary conditional

$$\frac{\begin{array}{c} P(B \rightarrow A) \geq r \\ P(B) \geq \rho \end{array}}{\therefore P(A) \geq r + \rho - 1} \tag{6}$$

These results are then reexpressed so as to feature the nearness of the probability to 1:

$$\frac{\begin{array}{c} P(B \rightarrow A) \geq 1 - \epsilon \\ P(B) \geq 1 - \epsilon \end{array}}{\therefore P(A) \geq 1 - 2\epsilon} \quad (7) \qquad \frac{\begin{array}{c} P(A \mid B) \geq 1 - \epsilon \\ P(B) \geq 1 - \epsilon \end{array}}{\therefore P(A) \geq (1 - \epsilon)^2} \quad (8)$$

Comparing these with necessary (verity) inference forms, Suppes observes that rules involving two premises (e.g., A, $A \rightarrow B \therefore B$) have corresponding probabilistic forms (e.g., (7)) in which the probability bound on the conclusion (of the necessary form) is less that that of the premises, though for single premise inferences (e.g., $A \therefore A \vee B$) this bound remains unchanged. Stated formally (*Suppes 1966*, 54):

Theorem 1. *If $P(A) \geq 1 - \epsilon$ and A logically implies B, then $P(B) \geq 1 - \epsilon$.*

Theorem 2. *If each of the premises $A_1, \ldots, A_n$ has probability at least $1 - \epsilon$ and these premises [conjointly] logically imply B, then $P(B) \geq 1 - n\epsilon$. Moreover, in general the lower bound $1 - n\epsilon$ cannot be improved on, i.e., equality holds in some cases whenever $1 - n\epsilon \geq 0$.*

Proofs of these results are quite straightforward except for the 'moreover' part of Theorem 2. This is the example which Suppes gives to show that

the lower bound for the conclusion in Theorem 2 can be attained (*1966*, 55):

> The example I use is most naturally thought of as a temporal sequence of events $A_1, \ldots, A_n$. Initially we assign
>
> $$P(A_1) = 1 - \epsilon$$
> $$P(\overline{A}_1) = \epsilon.$$
>
> Then [assigning]
>
> $$P(A_2 \mid A_1) = \frac{1 - 2\epsilon}{1 - \epsilon}$$
> $$P(A_1 \mid \overline{A}_1) = 1,$$
>
> and more generally
>
> $$P(A_n \mid A_{n-1}A_{n-2} \ldots A_1) = \frac{1 - n\epsilon}{1 - (n-1)\epsilon}$$
> $$P(A_n \mid A_{n-1}A_{n-2} \ldots \overline{A}_1) = 1$$
> $$\vdots$$
> $$P(A_n \mid \overline{A}_{n-1}\overline{A}_{n-2} \ldots \overline{A}_1) = 1,$$
>
> in other words for any combination of preceding events on trials 1 to $n-1$ the conditional probability of A_n is 1, except for the case $A_{n-1}A_{n-2} \ldots A_1$.

Suppes then shows by induction that $P(A_n A_{n-1} \ldots A_1) = 1 - n\epsilon$. Two comments are here in order. First, Suppes neglects to show that an assignment of probabilities such as he describes is a possible one. Secondly, a much simpler example can be given. Consider the probability 'space' which is the unit square with probability 'being' the area of a subregion. Assume $0 < \epsilon < \frac{1}{n}$ and let the square be divided so that there are n mutually exclusive subregions $\overline{A}_1, \ldots, \overline{A}_n$, each with probability equal to ϵ. Then

$$\begin{aligned} P(A_1 A_2 \ldots A_n) &= 1 - P(\overline{A}_1 \vee \overline{A}_2 \vee \cdots \vee \overline{A}_n) \\ &= 1 - \sum_{i=1}^{n} P(\overline{A}_i) \\ &= 1 - n\epsilon. \end{aligned}$$

Looking back 119 years to De Morgan's *Formal Logic*, chapter 10, (discussed in §2.2 above) and comparing De Morgan's examples with those of

Suppes' we note, not surprisingly, that advances have been made: in clarity with regard to the nature of logic, probability and their interrelations; in separation of logical from non-logical subject matter; in the introduction of inequality relations, and in formal correctness. Except for forms involving conditional probability Suppes limits himself, as did De Morgan, to inference forms in which the formula involved in the conclusion is a (verity) necessary consequence of those involved in the premises. Results are derived as in a mathematical theory and there is no semantic criterion as to what constitutes a valid inference. We would describe what Suppes presents as some applications of probability notions to ordinary (two-valued) logic, not a probability *logic*.

In connection with his Theorem 2 we have noted Suppes' interest in a bound being best possible ("cannot be improved on"). Apparently he was unaware that Boole had, long before, investigated the subject from a general viewpoint. This is the topic of our next section.

§3.4. Bounds on probability—early history

The determination of probability bounds is in its own right a topic of interest. We shall later see, in chapter 6, applications to estimation in statistics, to fault analysis in combinational (switching) circuits, and to network reliability. But it has additional significance for our study in that it plays a fundamental role in the construction of a probability logic—meaning by this a genuine logic and not simply an application of probability to (verity) logic. We shall see that the notion of *best possible probability bounds* which we arrive at is independent of any particular events or any particular probability space. This is analogous to the property of validity in verity logic, which is independent of any particular propositions or models.

Our historical account of probability bounds requires back-tracking to the nineteenth century as the topic was initiated and quite extensively developed by Boole, chiefly in his *Laws of Thought*. We begin the discussion by quoting the opening paragraphs from chapter XIX, Of Statistical Conditions (*1854*, 295):

> 1. By the term statistical conditions, I mean those conditions which must connect the numerical data of a problem in order that those data may be consistent with each other, and therefore such as statistical observations might actually have furnished. The determination of such conditions constitutes an important problem, the solution of which, to an extent sufficient at least for the requirements of this work, I purpose to undertake in the present chapter, regarding it partly as an independent object of speculation, but partly also as a necessary supplement to the theory of probabilities already in some degree exemplified.

He then goes on to state the nature of the connection:

> 2. There are innumerable instances, and one of the kind presented itself in the last chapter, Ex. 7, [or (2), in §2.7 above] in which the solution of a question in the theory of probabilities is finally dependent upon the solution of an algebraic equation of an elevated degree. In such cases the selection of the proper root must be determined by certain conditions, partly relating to the numerical values assigned in the data, partly to the due limitation of the element required. The discovery of such conditions may sometimes be effected by unaided reasoning. For instance, if there is a probability p of the occurrence of an event A, and a probability q of the concurrence of the said event A, and another event B, it is evident that we must have
>
> $$p \overset{=}{>} q.$$
>
> But for the general determination of such relations, a distinct method is required, and this we proceed to establish.

As this quotation indicates, Boole's interest in the topic arose from a specific need connected with his probability method, but he notes its importance for obtaining consistency conditions on probability data. (In later writings he replaces the phrase 'statistical conditions' by 'conditions of possible experience'.)

When Boole speaks of the "general determination of such relations" he is apparently referring to a uniform procedure which, when applied in a given instance, results in consistency conditions. Also involved, though not explicitly mentioned, is another type of generality: since the conditions are derived using only general principles of probability theory they hold good for any probability situation, i.e., putting it in a formal setting, they hold good for any probability algebra (or space). And, as these conditions are expressed in terms of inequations, they are directly connected with the question of bounds. Thus, for the simple example Boole mentions, if $P(A) = p$, then 0 and p are, respectively, lower and upper bounds on $P(AB)$ good no matter what A and B may be. As another example, if $P(A) = p$ and $P(B) = q$, then for the conjunction AB,

$$0 \leq P(AB) \leq \text{ minimum of } p \text{ and } q.$$

Thus 0 and $\min(p,q)$ are lower and upper bounds on $P(AB)$. We proceed to the details of Boole's procedures. We say procedure*s* for he has more than one.

The approach he uses in *Laws of Thought* is not based directly on the notion of probability but indirectly through the operator 'n', where '$n(x)$' stands for the number of elements in a class x. Relations among such values lead first to relations among relative frequencies by dividing through by $n(1)$, the number of elements in the universal class 1, and then to probabilities by passing to the limit. (Boole is quite vague about this transition from relative frequencies to probabilities.) In the interests of brevity we shall be referring chiefly to upper bounds or, as Boole calls them, 'major limits' of a class. Minor limits are obtained by subtraction from 1 of major limits of the complementary class.

Boole first considers the problem of obtaining major and minor limits of a (Boolean) class expression in terms of the number of elements in the classes represented by the class symbols present in the expression. Thus for the class expression 'xy' the major limit is 'the least of $n(x)$ and $n(y)$'; and for any 'constituent' (i.e., basic conjunction) the major limit is the least of the values for the individual conjuncts. For the general case of any class expression Boole replaces the class expression by an equivalent logical sum of constituents (i.e., by a disjunctive normal form). The problem is stated as a proposition (*1854*, 300):

> PROPOSITION II.
>
> 6. To determine the major numerical limit of a class expressed by a series [i.e., logical sum] of constituents of the symbols x, y, z, &c., the values of $n(x), n(y), n(z)$, &c., and $n(1)$, being given.

The solution is given as a rule (*1854*, 301):

> RULE.—Take one factor from each constituent, and prefix to it the symbol n, add the several terms or results thus formed together, rejecting all repetitions of the same term; the sum thus obtained will be a major limit of the expression, and the least of all such sums will be the major limit to be employed.

Thus as an example Boole lists for $n(xy + x(1-y)z)$ the sums

1. $n(x)$	4. $n(y) + n(x)$
2. $n(x) + n(1-y)$	5. $n(y) + n(1-y)$
3. $n(x) + n(z)$	6. $n(y) + n(z)$

One readily verifies the correctness of the rule as giving *an* upper bound. But note the absence of a claim for its being the best, i.e., narrowest that could be obtained—Boole's concluding sentence in the RULE simply states that the least of all such sums "will be the major limit to be employed [for purposes of his probability method]". An example showing that Boole's RULE does not yield the optimal bound first appeared in *Hailperin 1965*, §4.

Boole goes on to consider the most general situation stating it as a problem, somewhat unclearly, as follows (*1854*, 304):

> PROPOSITION IV
>
> Given the respective number of individuals composed in any classes s, t, &c., logically defined, to deduce a system of numerical limits of any other class w, also logically defined.

By 'logically defined' Boole means that there are (Boolean) equations

$$s = \phi(x, y, \ldots, z), \quad t = \psi(x, y, \ldots, z), \quad \ldots, \quad w = F(x, y, \ldots, z) \tag{1}$$

defining $s, t, \ldots, w$ in terms of a set of (unspecified) class symbols $x, y, \ldots, z$. By his general method in logic Boole deduces from a system such as (1) that

$$w = 1\,A + 0\,B + \frac{0}{0}\,C + \frac{1}{0}\,D, \tag{2}$$

where A, B, C, D are constituents on $s, t, \ldots$ determined by the method. The interpretation which Boole gives to this peculiar equation involves use of his indefinite class symbol 'v' but can be seen to be equivalent to

$$\begin{cases} A \subseteq w \subseteq A + C \\ D = 0. \end{cases} \tag{3}$$

Here $D = 0$ is the necessary condition that there be the (two-sided inclusion) solution for w given by the first equation. It is the case that (3) is always a correct consequence of (1). (See *Hailperin 1986*, §2.6.) For the simple example, introduced in §2.5, where

$$s = x, \quad t = xy, \quad w = y, \tag{4}$$

we have seen that Boole's technique gives

$$w = \frac{t}{s} = st + 0\,s\bar{t} + \frac{0}{0}\,\bar{s}\bar{t} + \frac{1}{0}\,\bar{s}t$$

with the meaning

$$\begin{cases} st \subseteq w \subseteq st + \bar{s}\bar{t} \\ \bar{s}t = 0. \end{cases}$$

Returning to the general case we state Boole's solution to the problem proposed in his PROPOSITION IV: the major limits of w are given by the major limits of $A + C$, these limits being expressed in terms of the given (numerical) values of $s, t, \ldots$. The equation $D = 0$, from which w is absent, provides the necessary condition for the existence of a solution, and implies that all minor limits of D be ≤ 0. Thus for the simple example (4) for which $A+C$ is $st+\bar{s}\bar{t}$, the major limits for w are, by Boole's RULE,

$$n(s) + n(\bar{t}), \quad n(t) + n(\bar{s}). \tag{5}$$

The minor limit of $\bar{s}t$, the D for this example, is $n(\bar{s}) + n(t) - 1$, (i.e., $1 -$ major limit of the complement $s + \bar{s}\bar{t}$) which gives

$$n(\bar{s}) + n(t) - 1 \leq 0,$$

or

$$n(t) \leq n(s), \tag{6}$$

as the necessary condition for a solution. (Clearly correct for $s = x$ and $t = xy$. Note that by adding $n(\bar{t})$ to both sides of (6) the first limit in (5) is not less that 1 and hence can be dropped.)

With regard to this method, later referred to by Boole as the 'prior' method, he has the disclaimer (*1854*, 310):

> 13. It is to be observed, that the method developed above does not always assign the narrowest limits which it is possible to determine. But in all cases, I believe, sufficiently limits the solutions of questions [of the general kind he has proposed] in the theory of probabilities.

No example is given to justify the assertion in the first sentence. But since the method depends on Boole's above cited RULE, the example in our *1965*, §4, referred to above, does provide one. Most interestingly, Boole then goes on to outline an entirely different method (*1854*, 310–11):

> The problem of the determination of the narrowest limits of numerical extension of a class is, however, always reducible to a purely algebraical form.* Thus, resuming the equation
>
> $$w = A + 0\,B + \frac{0}{0}C + \frac{1}{0}C,$$
>
> let the highest inferior numerical limit of w be represented by the formula $a\,n(s) + b\,n(t) \ldots + d\,n(1)$, wherein $a, b, c, \ldots, d$ are numerical constants to be determined, and s, t, &c., the logical symbols of which A, B, C, D are constituents. Then
>
> $$\begin{aligned} &a\,n(s) + b\,n(t) \ldots + d\,n(1) \\ &\qquad = \text{ minor limit of } A \text{ subject to the condition } D = 0. \end{aligned}$$

He then describes how to find the highest inferior limit:

> Hence if we develop the function
>
> $$as + bt \ldots + d,$$
>
> reject from the result all constituents which are found in D, the coefficients of those constitutents which remain, and are found also in A, ought not individually to exceed unity in value, and the coefficients of those constituents which remain, and which are not found in A, should individually not exceed 0 in value. Hence we shall have a series of inequalities of the form $f \overline{\overline{<}}\, 1$, and another series of the form $g \overline{\overline{<}}\, 0$, f and g being linear functions of a, b, c, &c. Then those values of $a, b..d$, which, while satisfying the above conditions, give to the function
>
> $$a\,n(s) + b\,n(t) \ldots + d\,n(1),$$
>
> its highest value must be determined, and the highest value in question will be the highest minor limit of w. To the above we may add the relations similarly formed for the determination of the relations among the given constants $n(s), n(t), \ldots\ n(1)$.

Boole's footnote (here omitted) refers to a lost manuscript of which he recollects only the impression that the principal of it was the same as just described, that it was developed in considerable detail, and its sufficiency was formed. (But see our *1965*, §7, where we express the opinion that his recollection concerning its sufficiency was in error.)

This sketch of a 'purely algebraic form' which Boole presents merits special mention: it is an early historical example of a linear programming problem, as well as a correct method of finding optimal probability bounds. However, aside from the obscurity due to his peculiar logical methods, there are three major gaps that need to be filled so as to make it cogent. One is a justification for using A, a sum of constituents on $s, t, \ldots$, in place of $F(x, y, z, \ldots)$. For, when the variables $s, t, \ldots$ in A are replaced by the functions of $x, y, z, \ldots$ as given in (1), the resulting expression, while implying $F(x, y, z, \ldots)$, need not be equivalent to it. (See our simple example (4) above). Another gap is the assumption that the best lower bound is a *linear* function of the given values $n(s), n(t), \ldots$. Finally, there is no proof that the bounds are *best* possible.

Assuming that the gaps are filled one can justify Boole's assertion that the best possible lower bound ('highest minor limits of w') is obtainable as described, namely by finding the maximum value of the linear form

$$a\,n(s) + b\,n(t) + \cdots + d\,n(1) \tag{7}$$

subject to the linear inequation system which he gives. In showing this Boole resorts to an expansion of the numerico-logical expression

$$as + bt + \cdots + d, \tag{8}$$

where, $a, b, \ldots, d$ are numerical variables and $s, t, \ldots$ are logical symbols. Such an expression, which seems queer to us, is a feature of his algebraic method of doing logic. However one can also obtain the result in a less mysterious manner. Let

$$\bigvee_s K_i, \quad \bigvee_t K_i, \quad \ldots, \quad \bigvee_1 K_i$$

be respectively the expansions of $s, t, \ldots, 1$ as sums of constituents on $s, t, \ldots$ (The expansion of 1 has all possible constituents on $s, t, \ldots$) Since the operator n distributes over a sum of mutually exclusive classes, we have

$$n(s) = \sum\nolimits_s n(K_i), \quad n(t) = \sum\nolimits_t n(K_i), \quad \ldots, \quad n(1) = \sum\nolimits_1 n(K_i).$$

By (i) multiplying these equations respectively by $a, b, \ldots, d$, (ii) adding to form a single equation, and (iii) collecting terms on the right, we obtain

$$a\,n(s) + b\,n(t) + \cdots + d\,n(1) = \sum_i h_i\,n(K_i), \tag{9}$$

where the h_i are linear combinations of $a, b, \ldots, d$. The indicated sum in (9) extends over all possible constituents K_i on $s, t, \ldots$. Since we seek the maximum value of (7) subject to the condition $D = 0$ we can delete any term $h_i\,n(K_i)$ in this sum if the K_i occurs in D—for if $D = 0$, then $n(K_i) = 0$ for any K_i in D. If (7) is to represent a lower bound for $n(A)$ subject to $D = 0$, then for each K_i which is present in A its coefficient h_i on the right in (9) must not exceed 1, and for each K_i not present in A, its coefficient h_i must not exceed 0. Hence if a lower bound for $n(A)$ is representable in the form (7) then its maximum value, subject to the described linear inequalities, is the best possible lower bound. (Boole says nothing about mathematical techniques for finding the maximum of a linear form under linear constraints—a subject of paramount interest in current linear programming theory.) To illustrate with a simple example what we have just described in general, let $s = x, t = xy$; we wish to obtain the highest minor limit of w where $w = y$. Here (see (4) above) $A = st$ and $D = \bar{s}t$. Set $L = a\,n(s) + b\,n(t) + c\,n(1)$, a linear form in variables a, b, c. Then

$$\begin{aligned} L = {} & a(n(st) + n(s\bar{t})) \\ & + b(n(st) + n(\bar{s}t)) \\ & + c(n(st) + n(s\bar{t}) + n(\bar{s}t) + n(\bar{s}\bar{t})) \end{aligned}$$

which then gives

$$\begin{aligned} L = {} & (a + b + c)n(st) \\ & + (a + 0 + c)n(s\bar{t}) \\ & + (0 + b + c)n(\bar{s}t) \\ & + (0 + 0 + c)n(\bar{s}\bar{t}). \end{aligned}$$

Since $\bar{s}t = D$ we delete the term $n(\bar{s}t)$ and, since st is the only constituent of A, we set

$$\begin{aligned} a + b + c &\le 1 \\ a + c &\le 0 \\ c &\le 0. \end{aligned} \tag{10}$$

The maximum of L subject to these constraints is (by linear programming theory) attained at a vertex of the polytope specified by (10). Here there is only one vertex, namely $(0, 1, 0)$. Hence $0\,n(s) + 1\,n(t) + 0\,n(1) = n(t)$ is the minor limit of w $(= y)$.

Perhaps the easiest and most informative way of filling the gaps mentioned above so as to make Boole's argument cogent, is by way of the duality theorem of linear programming. As we shall be referring to it in our next section it will be convenient to postpone completion of the discussion of Boole's 'purely algebraic form' until then.

In a subsequent paper appearing in the *Philosophical Magazine* Boole presents a distinctively different method of obtaining consistency conditions and bounds on the probability of a logical function. His opening paragraph declares it to be "an easy and general method of determining such conditions" and, referring to his earlier method in the *Laws of Thought*, "... the method there developed is difficult of application and I am not sure that it is equally general with the one I am about to explain."

In this paper (*1854*a) Boole drops the use of 'n' and expresses everything in terms of probability. The general problem proposed in PROPOSITION IV quoted above is expressed via the equations

$$P(\phi(x,y,\ldots,z)) = p, \quad P(\psi(x,y,\ldots,z)) = q, \quad \ldots,$$
$$P(F(x,y,\ldots,z)) = w \tag{11}$$

where $\phi, \psi, \ldots, F$ are given Boolean functions, with $\phi(x,y,\ldots,z)$, $\psi(x,y,\ldots,z), \ldots$ having respective probabilities $p, q, \ldots$. The object is to find bounds on w, the probability of $F(x,y,\ldots,z)$. Boole's first step is to convert equations (11) to a set of linear equations in the probabilities of the constituents on $x,y,\ldots,z$. This is accomplished by replacing each logical function by its expansion in terms of constituents on $x,y,\ldots,z$ and distributing P over the sums of constituents. If we denote constituent probabilities generically by k_i the resulting equations take the form

$$\sum\nolimits_\phi k_i = p, \quad \sum\nolimits_\psi k_i = q, \quad \ldots, \quad \sum\nolimits_F k_i = w. \tag{12}$$

To these equations Boole adjoins the normative conditions that the sum over all k_i be 1, and that for each i, $k_i \geq 0$. Finding the minimum value of the linear form $\sum_F k_i$ subject to these linear constraints is a linear programming problem. Boole took it for granted that the minimum value thus obtained would be 'best possible'. A mathematical proof may be found in our *1965*, §3. In effect Boole treats the problem as a *parametric* linear program and expresses his result in terms of the parameters $p, q, \ldots$ as follows.

First he shows how to successively eliminate variables one by one from a system of linear (equations and) inequations. Though he makes no mention of it, the method had been described earlier by Fourier in the 1820s. (See *Dantzig 1963*, 84) Next he applies (Fourier) elimination (of the k_i) to the combined system, i.e., to (12) plus the normative conditions, with w taken as a parameter along with $p, q, \ldots$. This ultimately results in a series of equations

$$\begin{array}{lll} L_1 \leq w, & L_2 \leq w, & \ldots \\ w \leq U_1, & w \leq U_2, & \ldots \end{array} \tag{13}$$

in which the L_i and U_i are linear forms in the parameters $p, q, \ldots$. It is clear that the L_i are minor limits, and the U_i major limits, of $P(F(x,y,\ldots,z))$. The largest of the L_i and the smallest of the U_i are what Boole takes as the narrowest limits on $P(F(x,y,\ldots,z))$ that are expressible in terms of $p, q, \ldots$. Finally, eliminating w from (13) results in a set of inequalities on $p, q, \ldots$ which are the necessary conditions that equations (11) have a solution for w.

When Boole describes this method as 'easy and general' as compared with the 'purely algebraic form' sketched in his *Laws of Thought* he underestimates the computational difficulties involved, for Fourier elimination can result in dauntingly large (exponentially exponential) numbers of inequalities.

In concluding this section we should like to mention two papers from the Russian literature—one from the 19th and the other from the 20th century—both presenting essentially the same (incomplete) solution that Boole gave to his general probability problem. The only significant respect in which they differ from Boole's is in the use of inclusion instead of Boole's indefinite class symbol to indicate a logical range of values. Boole's general problem has been discussed by us in our *1988b*, §9 (and in much fuller detail in our *1986*, chapters 4 and 5). The matter is related to the question of bounds in that Boole's solution often leads to a range of values as the answer. But this range need not be the same as the full range of possible values allowed by the given conditions of the problem, i.e., his method need not give a complete solution. This same failing applies to the two Russian papers which we now briefly discuss.

Poretzky's *1887* provides five worked examples of which four are taken without change from *Boole 1854*. These examples of Boole's have all been discussed in our *1986*. Poretzky's fifth example is of no interest here as it is a simple one with an exact value as a solution. *Čirkov 1971* has two worked examples of which the first has an exact value as a solution, but the second results in a range of values.

The problem is stated by Čirkov in technical applications-like form referring to 'automata', '0,1 signals', etc. However in purely mathematical terms it is as follows:

Given events A_1, A_2, A_3, A_4 of which it is only known that

$$P(A_1) = a_1, \quad P(A_2) = a_2, \quad P(A_3) = a_3, \quad P(\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4) = b,$$

find $P(A_4)$.

Letting r be any value for $P(A_4)$ which satisfies the hypothesis Čirkov finds that the bounds on r are given by

$$\frac{(1-b-a_1)(1-b-a_2)(1-b-a_3)}{(1-b)^2} \leq r \leq 1-b.$$

One can construct counter-examples that show that the lower limit is not best possible.[4] Application of the technique described below in §4.5

4. Take A_1, A_2, A_3 to be mutually exclusive and $P(A_1) = P(A_2) = P(A_3) = P(\overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4) = 1/4$. Then Čirkov's lower bound is $\frac{1/2\cdot1/2\cdot1/2}{(3/4)^2}$, which is greater than 0, a value $P(A_4)$ can take on.

yields the following optimal bounds:

$$\begin{aligned}\text{Lower } &= L = \max(0,\ 1-(a_1+a_2+a_3+b))\\ \text{Upper } &= U = \min(1,\ 1-b),\end{aligned}$$

with

$$0 \le L \le U \le 1$$

being the consistency condition on the parameters. Čirkov's problem is a slight generalization of one of Boole's (*1854*, 281) to which it reduces by setting $a_3 = 0$. The result then is the same as Boole's.

§3.5. Best possible probability bounds

After Boole's 1854 work on the probability of logical functions there seems to be nothing of note on this topic until Fréchet's *1935*. Although in this paper Fréchet cites Boole for a simple inequality, there is no mention of Boole's general results described in our preceding section.

Fréchet addresses himself to the following problem. What is the best that can be said about the probability P_a of an alternation of events $H_1 \vee H_2 \vee \cdots \vee H_n$ if all one knows is that the events have, respectively, probabilities $p_1, \ldots, p_n$? He readily shows that

$$p_1 + \cdots + p_n - (n-1) \le P_a \le p_1 + \cdots + p_n$$

and hence, setting

$$\begin{aligned}\omega(p_1, \ldots, p_n) &= \max\{0,\ p_1 + \cdots + p_n - (n-1)\}\\ \Pi(p_1, \ldots, p_n) &= \min\{1,\ p_1 + \cdots + p_n\},\end{aligned}$$

that

$$\omega(p_1, \ldots, p_n) \le P_a \le \Pi(p_1, \ldots, p_n). \tag{1}$$

Fréchet derives the inequalities (1) by simple direct means applied to the specific compound event involved. Boole on the other hand had, as we have seen in the preceding section, presented general methods for obtaining bounds on the probability of an arbitrary (logically expressed) compound event; moreover, there was the additional generality that the components

H_i of the compound event may be subject to additional, logically expressed, conditions. Though Boole declares that his method yields the 'narrowest limits' no justification of this is given, an essential for which is a criterion for judging when a bound is best possible. Fréchet's paper does implicitly provide one for his particular case, namely: the functions $\omega(p_1, \ldots, p_n)$ and $\Pi(p_1, \ldots, p_n)$ are *best possible bounds* for $P_a = P(H_1 \vee \cdots \vee H_n)$ if, for any pair of functions $f(p_1, \ldots, p_n)$ and $F(p_1, \ldots, p_n)$ such that for all possible values $p_1, \ldots, p_n$ in the unit interval and chance events $H_1, \ldots, H_n$ [in a probability space] for which $p_1, \ldots, p_n$ are their respective probabilities and

$$f(p_1, \ldots, p_n) \le P \le F(p_1, \ldots, p_n),$$

one has [for all $p_1, \ldots, p_n$ in the unit interval]

$$f(p_1, \ldots, p_n) \le \omega(p_1, \ldots, p_n) \le \Pi(p_1, \ldots, p_n) \le F(p_1, \ldots, p_n).$$

It is clear that a comparable criterion applies to any compound event and not just this particular one—provided that its components (the H_i) have unrestricted probabilities. This would not be the case if the H_i were subject to restrictions.

Fréchet employs his criterion to show that the functions ω and Π are the best that one can do. Considering the lower bound in (1), he asserts that for any set of reals $p_1, \ldots, p_n$ in the unit interval one can define a chance event G of probability $p' = \omega(p_1, \ldots, p_n)$, then events H_j with respective probabilities p_j ($j = 1, \ldots, n$) which take place only if G does, so that $P_a \le p'$. For these events

$$f(p_1, \ldots, p_n) \le P_a = p' = \omega(p_1, \ldots, p_n),$$

thus establishing the result for the lower bound.

Some 30 years later, in *Hailperin 1965*, attention was called to the three methods Boole had developed for obtaining probability bounds: (i) the 'prior' method of his chapter XIX in *1854*, (ii) the 'purely algebraic form' which he claimed did give narrowest limits, and (iii) the 'easy and general' method of his *1854a* of which he was unsure that it was equivalent to the purely algebraic form. Additionally, our *1965* gave an example to show that the 'prior' method did not in all cases provide a best possible bound; also, that methods (ii) and (iii) were primal and dual—hence equivalent—forms of the same linear program, and that the solution of the linear program did give the best possible bounds. Implicit in our paper, though not explicitly stated until *1976*, §5.6, was the conclusion that the solution to Boole's General Probability Problem consisted in specifying the best possible upper and lower bounds on the probability sought. It will be instructive to describe these ideas in the context of a specific worked example. The description will refer to the notion of a probability algebra, or space—as in our *1965*—rather than to our subsequently developed notion of a probability logic.

The simple example we select is the one already referred to in §2.5: given $P(A_1A_2) = p$, $P(A_2) = q$, find $P(A_1)$. Since, as is the case here, the given conditions need not determine a unique value for the unknown, we seek best possible functions of p and q which between them include any possible value for $P(A_1)$. The qualification 'best possible' is important since it is not sufficient simply to derive, by probability principles or axioms of probability algebras, a conclusion about $P(A_1)$. We wouldn't be sure that the conclusion was the most general result.

Let $\alpha = \alpha(p,q)$ be the set of all possible values of w such that there is a probability algebra $< \mathfrak{B}, P >$ with $A_1, A_2 \in \mathfrak{B}$, $P(A_1A_2) = p, P(A_2) = q$ and $w = P(A_1)$. If α is non-empty (it could be empty if, for example, q were less than p), then as a non-empty bounded set of real numbers it has a least upper bound $U(p,q)$. By virtue of its definition $U(p,q)$, as a function of p and q, is *the best possible upper bound function for* $P(A_1)$ *subject to the given conditions.* Similarly $L(p,q)$, defined as the greatest lower bound of α, is the best possible lower bound. But these definitions are non-effective, furnishing no means for computing values of $U(p,q)$ and $L(p,q)$. To obtain computable forms we approach the problem from another direction.

Suppose we have a probability algebra $< \mathfrak{B}, P >$ with $A_1, A_2 \in \mathfrak{B}$. Set $k_1 = P(A_1A_2)$, $k_2 = P(A_1\overline{A}_2)$, $k_3 = P(\overline{A}_1A_2)$, $k_4 = P(\overline{A}_1\overline{A}_2)$, so that $P(A_1) = k_1 + k_2 (= w)$. As Boole does in his 'easy and general' method we express the given conditions in the form

$$\begin{aligned} 1k_1 + 0k_2 + 0k_3 + 0k_4 &= p \\ 1k_1 + 0k_2 + 1k_3 + 0k_4 &= q \\ 1k_1 + 1k_2 + 1k_3 + 1k_4 &= 1 \\ k_1, k_2, k_3, k_4, &\geq 0, \end{aligned} \tag{2}$$

where the first two equations are the explicit conditions, and the remaining lines are implicit ones. In matrix notation these conditions become

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ k_3 \\ k_4 \end{bmatrix} = \begin{bmatrix} p \\ q \\ 1 \end{bmatrix} \tag{3}$$

$$[k_1 \quad k_2 \quad k_3 \quad k_4]^{\mathrm{T}} \geq \mathbf{0}.$$

Now define $\beta(p,q)$ to be the set of values $w = k_1 + k_2$ where (k_1, k_2, k_3, k_4) is any quadruple which satisfies the system (3), considered simply as an

algebraic system. One can show that the sets $\alpha(p,q)$ and $\beta(p,q)$ are the same and that $U(p,q)$ and $L(p,q)$ are its maximum and minimum. Hence finding best possible bounds for $P(A_1)$ is equivalent to solving the linear programming problem: find the max (and min) of the linear form $k_1 + k_2$ subject to the linear constraints (3). If, following Boole, we apply Fourier elimination (of k_1, k_2, k_3, k_4) to the system we obtain formed by adjoining the equation $w = k_1 + k_2$ to (3), we obtain

$$\begin{gathered} 0 \leq q \\ q \leq p \leq 1 \\ q \leq w \\ w \leq q + 1 - p \\ 0 \leq w, \ w \leq 1 \end{gathered} \tag{4}$$

so that from the last four inequations we have,

$$\begin{aligned} \max w &= \min\{q + 1 - p, 1\} \\ \min w &= \max\{0, q\}, \end{aligned} \tag{5}$$

and from the first two the consistency condition,

$$0 \leq q \leq p \leq 1, \tag{6}$$

for the existence of the solution.

By virtue of the duality theorem of linear programming an equivalent problem to finding the minimum of $(w =)\ 1k_1 + 1k_2 + 0k_3 + 0k_4$, subject to (3), is finding the maximum of $(u =)\ px_1 + qx_2 + 1x_3$, subject to

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \tag{7}$$

with the x_i unrestricted. Note that the coefficients in the objective function w of the primal form, i.e., $1, 1, 0, 0$, now are the constant terms on the right side of (7), that the constant terms on the right side of the equations in (3), i.e., $p, q, 1$, now are the coefficients of the objective function u of the dual form, and that the matrix of 0s and 1s in (7) making up the coefficients of the x_i, is the transpose of the matrix of coefficients in (3). This dual form of the linear programming problem has the significant feature that the parameters p, q now are no longer involved in specifying the region of feasible solutions, and hence the polytope specified by the inequalitites

(7) is expressed in purely numerical terms. The theory tells us that the optimal value of the objective function is obtained at a vertex, i.e., at a basic feasible solution $(x_1{}^{(i)}, x_2{}^{(i)}, x_3{}^{(i)})$. For our example these corner points are $(0,0,0)$ and $(0,1,0)$, so that the maximum value for u is the maximum of $p0+q0+0$ and $p0+q1+0$, that is, $\max\{0, q\}$.

We return to the unfinished business of §3.4 concerning Boole's use of $A(s,t,\ldots)$ in place of $F(x,y,\ldots)$ to obtain the optimal lower bound. We use our ongoing simple example to illustrate the ideas. Set $s = xy$, $t = y$, and $w = x$. Then Boole's solution for w in terms of s and t is (using inclusion instead of Boole's indefinite symbol):

$$\begin{cases} st \subseteq w \subseteq st + \bar{s}\bar{t} \\ \bar{s}t = 0, \end{cases}$$

and the method entails obtaining the optimum value for $P(st)$ so as to provide a lower bound for $P(w)$. (In this example $A(s,t,\ldots)$ is st, and $F(x,y,\ldots)$ is $W[=x]$.) In the table below we have, as a brief study will show, a combined representation of the relationships between x, y and s, t, w as well as the conditions expressed by (7).

$K_i(x,y)$	$s(=xy)$	$t(=y)$	1		st		$w(=x)$	
xy	1	1	1		1		1	
$x\bar{y}$	0	0	1	$(\le)$	0	$(\le)$	1	(Table 1)
$\bar{x}y$	0	1	1		0		0	
$\bar{x}\bar{y}$	0	0	1		0		0	

In this table the '1' or '0' in a column indicates the presence or absence of the constituent (of column 1) in the expansion of the term (as a function of x, y) heading the column. (Lines 2 and 4 of the table show that w can't be a truth-function of s and t since it is assigned different values for the same s and t values.) Ignoring columns 1 and 5 gives us the system (7), with $s, t, 1$ playing the role of x_1, x_2, x_3. Note, in (7), that any set of values (x_1, x_2, x_3) making its line 4 true also makes its line 2 true since $0x_1 + 0x_2 + 1x_3 \le 0$ implies $0x_1 + 0x_2 + 1x_3 \le 1$. Thus if in Table 1 we replace, under w, the '1' in line 2 by '0'—which then becomes the same as column 5—we lose nothing from the set of feasible points containing the optimal value. Thus the optimal value for $P(st)$ will be the same as that for $P(w)$. For the

general case there will be a table similar to Table 1:

$K_i(x,y,\ldots)$	$s\ t\ \ .\ .\ .\ u$	1		w
K_1		1		w_1
.	(1's and 0's)	1		w_2
.		.	$(\le)$	.
.		.		.
.		.		.
.		.		.
K_{2^n}		1		w_{2^n}

If for any pair $K_i, K_j, (i \neq j)$ the values under $s, t, \ldots, u$ are the same in both rows i and j but there are different values in these rows under w, then the value under w which is '1' can be replace by '0' without changing the feasible set (since ≤ 0 implies ≤ 1). We carry out this operation until no change is possible and let the resulting column be headed by 'w^*'. But these values under w^* are the same as that for a function of $s, t, \ldots, u$ obtained by taking the logical sum of constituents on $s, t, \ldots, u$ that is maximal with respect to implying w, and this is Boole's $A(s,t,\ldots,u)$.[5]

Note that the described procedure reduces the number of distinct rows in the matrix inequality and hence, after duplicate rows are deleted, reduces the amount of computation needed to find the vertices of the polytope.

The linear programming approach to logico-probability questions pioneered by Boole can be used to derive a large variety of probability identities, as was noted by Rényi in his *1970*, §2.6. Rényi considers linear relations of the form

$$\sum_{i=1}^{N} c_i P(F_i) = 0 \quad (\text{or, also, } \le 0, \ge 0) \tag{8}$$

where the c_i are real constants and the $F_i = F_i(A_1, \ldots, A_n)$ are Boolean polynomials of the events $A_1, \ldots, A_n$. By direct argument Rényi shows (Theorem 2.61, p. 64) that (8) is an identity, i.e., is valid in all probability spaces $S = (\Omega, \mathcal{A}, P)$, if it is valid in the 2^n special cases when each of the $A_1, \ldots, A_n$ is allowed to be either 0 or Ω, that is, if it is valid in the trivial space with just the two events 0 and Ω. After proving the theorem Rényi remarks (p. 68 with a reference to *Ha[i]lperin 1965*) that one may also view the result from the viewpoint of linear programming, but says nothing further. It is a rather nice application of the technique. We shall be presenting it in chapter 6.

5. Proof is given in (i) on page 159 of *Hailperin 1986*.

After our *1965*, four papers appeared which were concerned with logical-probability or probability-logical questions and which used linear programming. Apparently all four were independent developments since none of them mentions earlier or contemporaneous sources. Two of these will be described in chapter 6. The other two, which have logic as their primary interest, are discussed in the next two sections.

§3.6. Transmission of uncertainties in inferences

Featured in *Adams-Levine 1975* is an investigation of how uncertainties are transmitted from premises to conclusion in deductive inferences. This notion of inference is a generalized one and need not be necessary, i.e., deductively sound: any ordered pair $<< \phi_1, \ldots, \phi_n >, \psi >$ consisting of an n-tuple $< \phi_1, \ldots, \phi_n >$ of sentences (called the 'premises') and a sentence ψ (called the 'conclusion') is referred to as an *inference.* The *uncertainty* of a sentence is (by definition) the probability of its denial. The central question considered is:

If the premises ϕ_i of an inference have uncertainties that, respectively, do not exceed ϵ_i $(i = 1, \ldots, n)$ what is the maximum uncertainty of the conclusion ψ?

Since $P(\overline{\phi}) \leq \epsilon$ is equivalent to $\epsilon \leq P(\phi) \leq 1$ the notion of uncertainty of a sentence not exceeding ϵ is equivalent to that of its being within ϵ of 1. We have noted (in §3.3 above) that Suppes had presented some inferences featuring premises with probabilities near 1. What is different here is that with Adams-Levine the conclusion formula need not be a necessary consequence of the premise formulas. Moreover, they address themselves to a general class of inferences and not just specific inferences.

Their paper states that the sentences can be from an arbitrary first-order language. But many logical notions that are appealed to (e.g., logical implication, logical equivalence, consistency) would in general be non-effective if applied to first-order languages; and, moreover, nowhere is any use made of quantifier structure. Accordingly we shall assume that only the sentential (propositional, truth-functional) structure of sentences is relevant. With this understanding we see that the Adams-Levine question is a special subcase of Boole's 'general' probability problem, as extended to inequalities in

Hailperin 1965 §6, namely:

$$\begin{aligned} &\textit{Given:} \quad P(\bar{\phi}_i) \leq \epsilon_i, \quad (i = 1, \ldots, n) \\ &\textit{find:} \quad \text{the best possible upper bound for } P(\bar{\psi}) \end{aligned}$$

(the bound to be in terms of $\epsilon_1, \ldots, \epsilon_n$). The method which Adams-Levine use to solve this problem is quite different from the (extended) Boole approach and merits discussion.

Though the inferences Adams-Levine consider are not required to be deductively sound, nor need the premises and conclusion be related, the method is unnecessarily encumbered (we believe) with logical distinctions such as whether the premises are consistent or inconsistent,[6] which subsets of the premises imply or do not imply the conclusion, and others. These distinctions, to be sure, are involved later on in the paper, for example in tracing the effects of particular subsets of the premises on the conlusion's uncertainty. But, as we shall see, the maximum uncertainty of the conclusion is obtainable from a linear program which can be set up without appeal to these logical distinctions; and once such a linear program has been set up one can read off from the algebraic form the various logical items which the Adams-Levine method would have needed as prerequisite to its employment. An example or two will make this clear. But first an outline of the paper's procedure.

An initial step is the replacement of the conclusion ψ by another sentence defined in terms of the premises (the sentence may be ψ itself). The sentence replacing ψ for purposes of finding maximum uncertainty, which can take on either of two equivalent forms, me(I) or ms(I), is obtained as follows.

A subset P' of the set of premises P of an inference $I =< \phi_1, \ldots, \phi_n, \psi >$ is said to be *essential* for I if the complimentary set $P - P'$ does not imply ψ. An essential (for I) set is minimal if no proper subset is essential. Let $E_1, \ldots, E_s$ be minimal essential subsets for I. Then, where $\bigvee E_i$ is the disjunction of members of E_i (or the logically false sentence if E_i is empty), me(I) is defined to be $(\bigvee E_1) \wedge (\bigvee E_2) \wedge \cdots \wedge (\bigvee E_s)$ (or the logically true sentence if there are no minimal essential sets). A premise subset P' of P is *sufficient* (for I) if it implies ψ, and is minimal if no proper subset of it does. If the minimal sufficient subsets are $S_1, \ldots, S_r$ then ms(I) is defined to be $(\bigwedge S_1) \vee (\bigwedge S_2) \vee \cdots \vee (\bigwedge S_r)$ where $\bigwedge S_i$ is the conjunction of elements of S_i (or the logically true sentence if S_i is empty). If there are no minimal sufficient subsets then ms(I) is defined to be the logically false sentence. Thus, as an example, the inference

$$I =< A, B, -(AB), A \leftrightarrow -B >,$$

6. Inconsistent premises, e.g., ϕ and $\bar{\phi}$, need not be excluded—one can have $P(\bar{\phi}) \leq \epsilon_1$ and $P(\bar{\bar{\phi}}) \leq \epsilon_2$ so long as $1 \leq \epsilon_1 + \epsilon_2 \leq 2$.

with conclusion $A \leftrightarrow -B$, has the two minimal essential subsets

$$E_1 = \{A, B\} \text{ and } E_2 = \{-(AB)\}.$$

Hence

$$\mathrm{me}(I) = (A \vee B) \wedge (-(AB)) = A\bar{B} \vee \bar{A}B,$$

which happens in this case to be equivalent to the conclusion.

It is asserted in the paper that $\mathrm{me}(I)$ and $\mathrm{ms}(I)$ are equivalent and that either can replace ψ for the purpose of finding the maximum uncertainty of ψ. Proving the first of these assertions is straightforward, but not the second. It is further stated (p. 433):

> The reduction [replacement of ψ by $\mathrm{ms}(I)$] can be carried one step farther. If the premises are consistent then each premise can be replaced by a distinct atomic letter with the same replacements being made in $\mathrm{ms}(I)$ or $\mathrm{me}(I)$, and the new inference will be equivalent to the original with respect to uncertainty maximization. The same reduction can also be carried out when the premises are inconsistent, except that in this case it is necessary to add non-logical axioms to the language which specify in effect that sets of atomic formulas which correspond to inconsistent premise sets are inconsistent.

Note the correspondence of Adams-Levine's introducing atomic letters with Boole's introducing the letters $s, t, \ldots, u$ in place of the given functions of $x, y, \ldots, z$ (§§2.5 and 3.4 above). We now look at Adams-Levine's method of obtaining the linear inquality systems expressing $P(\bar{\phi}_i) \leq \epsilon_i$, and the linear form for $P(\overline{ms(I)})$. This requires examining the premises and determining the minimal essential subsets and the minimal negatively sufficient subsets of the premises. The first of these notions we have already introduced. As for the second , a subset of an inconsistent set of premises is *negatively sufficient* if its complement (in the set of premises) is consistent and sufficient (to imply the conclusion); if the premises are consistent, the sets are empty. Using these notions Adams-Levine construct a *minimal falsification matrix* from which *minimal falsification functions* are obtained to provide the linear forms in terms of which the linear program is stated. We shall not state the general definition of these two concepts but they may be inferred from a simple example we take from *Adams-Levine 1975* p. 435 (in this example $\mathrm{ms}(I)$ is equivalent to ψ):

> A simple illustration is the inference of the conclusion '$A \leftrightarrow -B$' from the three premises 'A', 'B', and '$-(A\&B)$' (the premises are both redundant and inconsistent). In this example there are two minimal sufficient premise sets, $S_1 = \{A, -(A\&B)\}$ and $S_2 = \{B, -(A\&B)\}$, two minimal essential premise sets $E_1 = \{A, B\}$ and $E_2 = \{-(A\&B)\}$, and two negatively sufficient premise sets, $NS_1 = \{A\}$, and $NS_2 = \{B\}$.

For this example (call it EXAMPLE 1) their minimal falsification matrix is:

$$\begin{array}{l|cccc}
 & & \textit{premises:} & & \textit{conclusion}: \\
 & A & B & -(A\&B) & A \leftrightarrow -B \\
\hline
E_1 = \{A, B\} & 1 & 1 & 0 & 1 \\
E_2 = \{-(A\&B)\} & 0 & 0 & 1 & 1 \\
NS_1 = \{A\} & 1 & 0 & 0 & 0 \\
NS_2 = \{B\} & 0 & 1 & 0 & 0
\end{array} \qquad (1)$$

In this matrix the rows are values of minimal falsification functions (one for each row) and are weighted—in the case of essential rows (here the first two) by non-negative reals $p_1, \ldots, p_s$ (in the example $s = 2$). and for the negatively sufficient rows (here the last two) by non-negative reals $q_1, \ldots, q_t$ (in the example $t = 2$) and the sum of all the weights is set equal to 1. Probabilities are then assigned by summing weighted *column* entries. Thus:

$$\begin{aligned}
P(\bar{A}) &= 1 \cdot p_1 + 0 \cdot p_2 + 1 \cdot q_1 + 0 \cdot q_2 \\
P(\bar{B}) &= 1 \cdot p_1 + 0 \cdot p_2 + 0 \cdot q_1 + 1 \cdot q_2 \\
P(\overline{AB}) &= 0 \cdot p_1 + 1 \cdot p_2 + 0 \cdot q_1 + 0 \cdot q_2 \\
\text{and} \qquad P(\overline{A \leftrightarrow \bar{B}}) &= 1 \cdot p_1 + 1 \cdot p_2.
\end{aligned}$$

This yields the linear program:

$$\text{maximize } p_1 + p_2$$

subject to

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ q_1 \\ q_2 \end{bmatrix} \leq \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

$$p_1 + p_2 + q_1 + q_2 = 1, \quad p_1, p_2, q_1, q_2 \geq 0.$$

We now show how Adams-Levine's minimal falsification matrix (1) can be obtained by a straightforward (to us, less obscure) method, Boole's 'purely algebraic form' described above in §3.4.

Introduce *letters* $\phi_1, \phi_2, \phi_3, \psi$ to represent the logical functions, i.e., set

$$\phi_1 \leftrightarrow A,\ \phi_2 \leftrightarrow B,\ \phi_3 \leftrightarrow \overline{AB},\ \text{and}\ \psi \leftrightarrow A\bar{B} \vee \bar{A}B. \tag{2}$$

Eliminate A and B from these equivalences by any one of a number of means, e.g., by computing all 2^3 basic conjunctions on ϕ_1, ϕ_2, ϕ_3, or by Boole's algebraic techniques (mentioned in §2.5 and §3.4 above). Then one obtains

$$\begin{gathered} \phi_1\phi_2\overline{\phi}_3 \vee \overline{\phi}_1\overline{\phi}_2\phi_3 \leftrightarrow \overline{\psi} \\ \overline{\phi}_1\overline{\phi}_2\overline{\phi}_3 \vee \overline{\phi}_1\phi_2\overline{\phi}_3 \vee \phi_1\overline{\phi}_2\overline{\phi}_3 \vee \phi_1\phi_2\phi_3 \leftrightarrow f, \end{gathered} \tag{3}$$

where f is the logically false sentence. (In general, the first line could be a two-sided inclusion of $\overline{\psi}$ by sums of basic conjunctions and not an equivalence.) As Boole would have put it, the basic conjunctions in the second line (corresponding to his $D = 0$ equation) are those combinations which are 'impossible' or excluded by the data. It is convenient to display this information in a table. Here $\frac{1}{0}$ stands for 'impossible' and we enter the table with values of $\overline{\phi}_i$ and $\overline{\psi}$ so as to match up with Adams-Levine's scheme.

	$\overline{\phi}_1$	$\overline{\phi}_2$	$\overline{\phi}_3$	$K_i(\overline{\phi}_1, \overline{\phi}_2, \overline{\phi}_3)$	$\overline{\psi}$
$i = 1.$	1	1	1	$\overline{\phi}_1\overline{\phi}_2\overline{\phi}_3$	$\frac{1}{0}$
2.	1	1	0	$\overline{\phi}_1\overline{\phi}_2\phi_3$	1
3.	1	0	1	$\overline{\phi}_1\phi_2\overline{\phi}_3$	$\frac{1}{0}$
4.	1	0	0	$\overline{\phi}_1\phi_2\phi_3$	0
5.	0	1	1	$\phi_1\overline{\phi}_2\overline{\phi}_3$	$\frac{1}{0}$
6.	0	1	0	$\phi_1\overline{\phi}_2\phi_3$	0
7.	0	0	1	$\phi_1\phi_2\overline{\phi}_3$	1
8.	0	0	0	$\phi_1\phi_2\phi_3$	$\frac{1}{0}$

(Table A)

Deleting rows containing $\frac{1}{0}$ (unrealizable, hence contributing no probability) produces

	$\overline{\phi}_1$	$\overline{\phi}_2$	$\overline{\phi}_3$	$K_i(\overline{\phi}_1, \overline{\phi}_2, \overline{\phi}_3)$	$\overline{\psi}$
$i = 2.$	1	1	0	$\overline{\phi}_1\overline{\phi}_2\phi_3$	1
4.	1	0	0	$\overline{\phi}_1\phi_2\phi_3$	0
6.	0	1	0	$\phi_1\overline{\phi}_2\phi_3$	0
7.	0	0	1	$\phi_1\phi_2\overline{\phi}_3$	1

(Table B)

This is identical with Adams-Levine's minimal falsification matrix (1) except that our columns are headed by the negated letters, and there is a different arrangement of the rows. Looking at these rows for which there is a 1 under $\overline{\psi}$ enables us to pick out the essential sets of premises (in the example they happen also to be minimal) and for the rows for which there is a 0 under $\overline{\psi}$ we obtain the negatively sufficient sets of premises. From Table B we see that the entire probability space consists of the four constituents listed under $K_i(\overline{\phi}_1, \overline{\phi}_2, \overline{\phi}_3)$. Assigning probabilities ('weights') p_1, p_2, q_1, q_2 to these constituents (letters 'p' for those corresponding to essential premise sets, and letters 'q' for those corresponding to negatively sufficient ones) enables us to write

$$P(\overline{A}) = p(\overline{\phi}_1) = P(\overline{\phi}_1\overline{\phi}_2\overline{\phi}_3 \vee \overline{\phi}_1\phi_2\phi_3) = 1p_1 + 0p_2 + 1q_1 + 0q_2,$$

and similarly for the other formulas of the inference, which reproduces the Adams-Levine linear program for the inference.

Since this EXAMPLE 1 doesn't show all the features of the Adams-Levine approach we present another one, one in which the premises are consistent and sufficient. This is taken from *Adams-Levine 1975* p. 44. Call it EXAMPLE 2. The premises are A, B, C and the conclusion, $A(B \vee C)$, is equivalent to ms(I). Thus we wish to find the maximum of $P(\overline{A} \vee \overline{BC})$ subject to $P(\overline{A}) \leq \epsilon_1$, $P(\overline{B}) \leq \epsilon_2$, $P(\overline{C}) \leq \epsilon_3$. In this case $\phi_1 \leftrightarrow A$, $\phi_2 \leftrightarrow B$, $\phi_3 \leftrightarrow C$ and there is no need to introduce the ϕ_i letters. Moreover there are no excluded combinations implied by the premises. Hence

	$\overline{A}$	$\overline{B}$	$\overline{C}$	$K_i(\overline{A}, \overline{B}, \overline{C})$	$\overline{A} \vee \overline{BC}$
$i = 1.$	1	1	1	$\overline{A}\,\overline{B}\,\overline{C}$	1
2.	1	1	0	$\overline{A}\,\overline{B}\,C$	1
3.	1	0	1	$\overline{A}\,B\,\overline{C}$	1
4.	1	0	0	$\overline{A}\,B\,C$	1
5.	0	1	1	$A\,\overline{B}\,\overline{C}$	1
6.	0	1	0	$A\,\overline{B}\,C$	0
7.	0	0	1	$A\,B\,\overline{C}$	0
8.	0	0	0	$A\,B\,C$	0

(Table C)

Lines 1–5 indicate which premise sets are the essential ones and lines 6–8 the negatively sufficient ones. Here there are minimal sets, line 4 and line 5 for the essential ones, and line 8 for the negatively sufficient ones. Thus we obtain Adams-Levine's minimal falsification matrix for this problem, namely

	$\overline{A}$	$\overline{B}$	$\overline{C}$	$\overline{A} \vee \overline{BC}$
$E_1 = \{A\}$	1	0	0	1
$E_2 = \{B, C\}$	0	1	1	1
$NS_1 = \{\}$	0	0	0	0

An equivalent reduction of Table C can be obtained without appeal to any of these special notions. We expand all functions as logical sums of basic conjunction on A, B, C and use the probabilities of the conjunctions as variables. Then the constraint system for the linear program is

$$\begin{array}{cccccccccl} k_1 & k_2 & k_3 & k_4 & k_5 & k_6 & k_7 & k_8 & & \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \leq \epsilon_1 & \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & \leq \epsilon_2 & \qquad \text{(Table D)} \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & \leq \epsilon_3 & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & = 1 & \end{array}$$

(The detached coefficient format for the system is used to conserve space.) It can be readily seen that columns headed by k_1, k_2, k_3, k_6, and k_7 can be dropped and one has an equivalent system. Even more readily can this be seen from the dual system, namely:

$$\begin{array}{cccccl} & w_1 & w_2 & w_3 & v & \\ 1. & 1 & 1 & 1 & 1 & \geq 1 \\ 2. & 1 & 1 & 0 & 1 & \geq 1 \\ 3. & 1 & 0 & 1 & 1 & \geq 1 \\ 4. & 1 & 0 & 0 & 1 & \geq 1 \\ 5. & 0 & 1 & 1 & 1 & \geq 1 \\ 6. & 0 & 1 & 0 & 1 & \geq 0 \\ 7. & 0 & 0 & 1 & 1 & \geq 0 \\ 8. & 0 & 0 & 0 & 1 & \geq 0 \end{array} \qquad \text{(Table E)}$$

$$w_1, w_2, w_3 \geq 0, \ v \text{ unrestricted}$$

We observe here that, since for each i, $w_i \geq 0$, any set of values w_1, w_2, w_3, v making the inequality in line 4 true also makes those in lines 1, 2, and 3 true. Thus these three lines—corresponding to non-minimal essential sets of premises—are (algebraically) redundant and can be deleted. Similarly line 8 implies lines 6 and 7 and they too can be deleted.

As in *Hailperin 1965*, §4, Adams-Levine also go over to a dual formulation so as to have, as illustrated by Table E, a constraint system which is independent of parameters (here the ϵ_i). These constraints can then be treated purely numerically. The parameters ϵ_i then appear only in the linear objective function—here $w_1\epsilon_1 + w_2\epsilon_2 + w_3\epsilon_3 + v$—which is to be minimized.

We continue with the second of the two independently conceived papers applying linear programming methods to logico-probability questions.

§3.7. Nilsson's probabilistic logic

Unlike the paper discussed in our preceding section, which views the topic as an application of probability to ordinary logic, Nilsson lays claim to a 'probabilistic logic' (*1986*, 72):

> In this paper we present a semantical generalization of ordinary first-order logic in which the truth values of sentences can range between 0 and 1. The truth value of a sentence in *probabilistic logic* is taken to be the *probability* of that sentence in ordinary first-order logic. We make precise the notion of the probability of a sentence through a possible-worlds analysis. Our generalization applies to any logical system for which the consistency of a finite set of sentences can be established.

We single out three items to be noted in this quotation:

(i) that a semantic generalization of ordinary first-order logic is presented,
(ii) that the notion of the probability of a sentence is made precise through a possible-worlds analysis and
(iii) that the generalization applies to any logical system for which the consistency of a finite set of sentences can be established.

We begin our discussion with (ii), making the notion of the probability of a sentence precise. Nilsson's analysis of "the probability of a sentence" starts out with (p. 72):

> To define what we mean by the *probability of a sentence* we must start with a sample space over which to define probabilities (as is customary in probability theory). A sentence S can be either *true* or *false*. If we were concerned about just the one sentence S, we could imagine two sets of *possible-worlds*—one, say $\mathcal{W}_1$, containing worlds in which S was *true* and one, say $\mathcal{W}_2$, containing worlds in which S was *false*. The actual world, the world we are actually in, must be in one of these two sets, but we might not know which one. We can model our uncertainty about the actual world by imagining that it is in $\mathcal{W}_1$ with probability p_1 and is in $\mathcal{W}_2$ with some probability $p_2 = 1 - p_1$. In this sense we can say that the probability of S (being true) is p_1.

Nilsson believes that he has to have a sample space whose subsets are assigned probabilities, and chooses for this space the set of all possible worlds. Then "the sentence S has probability p_1" is represented by

$$p(R \in \mathcal{W}_1(S)) = p_1, \qquad (1)$$

where R is the real world and $\mathcal{W}_1(S)$ is the subset of all possible worlds in which S is true. ('$\mathcal{W}_1(S)$' is used rather than '$\mathcal{W}_1$' to indicate its dependence on S.) We do not see the advantage in modelling S's uncertainty in this way. Why not simply

$$p(S) = p_1? \tag{2}$$

This would serve the same purpose, requiring only readjustment in one's thinking so as to be able to attribute probability to sentences,[7] and not necessarily to sets. The gain with (2) over (1) is the elimination of the ontological notions of 'real world' and 'set of possible worlds'.

Nilsson also introduces 'impossible worlds' (*1986*, 72):

> If we have L sentences, we might have as many as 2^L sets of possible worlds. Typically though, we will have fewer than the maximum number because some combinations of *true* and *false* values for our L sentences will be logically inconsistent. We cannot, for example, imagine a world in which S_1 is *false*, S_2 is *true* and $S_1 \wedge S_2$ is *true*. That is, some of the sets of the 2^L worlds might contain only impossible worlds.

He illustrates this with an example of three sentences $\{P,\ P \supset Q,\ Q\}$:

> The consistent sets of truth values for these three sentences are given by the columns in the following table:
>
> $$\begin{array}{rcccc} P & true & true & false & false \\ P \supset Q & true & false & true & true \\ Q & true & false & true & false \end{array} \tag{3}$$
>
> In this case, there are four sets of possible worlds each one corresponding to one of these four sets of truth values.

These sets of consistent truth value assignments—which 'correspond to' sets of possible worlds—are then taken as elements of a sample space on which to define a probability distribution. But for this there is no need for the possible-impossible worlds notion—one can obtain these four consistent assignments to which probability is attached just by an examination of the syntactic structure of the sentences (assuming, as Nilsson does, that the sentences are expressed as logical functions of a set of atomic sentences). Equivalently, and somewhat more conveniently, one can use constituents (basic conjunctions) on $S_1 = P,\ S_2 = P \supset Q,\ S_3 = Q$, that is *sentences*, to represent the consistent truth value assignments (the four columns), namely the four sentences

$$S_1S_2S_3,\ S_1\overline{S}_2\overline{S}_3,\ \overline{S}_1S_2S_3,\ \overline{S}_1S_2\overline{S}_3. \tag{4}$$

These are obtained by writing down all possible constituents on $S_1,\ S_2,\ S_3$ and deleting those which become inconsistent when S_1 is replaced by P, S_2 by $P \supset Q$ and S_3 by Q. The information contained in (3) is reproduced by an 'incidence matrix':

$$\begin{array}{rcccc} & S_1S_2S_3 & S_1\overline{S}_2\overline{S}_3 & \overline{S}_1S_2S_3 & \overline{S}_1S_2\overline{S}_3 \\ S_1(=P) & 1 & 1 & 0 & 0 \\ S_2(=P \supset Q) & 1 & 0 & 1 & 1 \\ S_3(=Q) & 1 & 0 & 1 & 0 \end{array} \tag{5}$$

where a 1 in the body of the table indicates that the formula to the left in that row appears unnegated, and a 0 that it appears negated, in the basic conjunction at the head of the column. Table (5) establishes a one-to-one correspondence between consistent assignments of truth values to $S_1,\ S_2,\ S_3$ and constituents on these letters. We can reproduce (5) in a slightly different way which serves to bring out the connection with Adams-Levine (§3.6) and Boole (§2.5). Consider the truth table on sentence variables $S_1,\ S_2,\ S_3$:

$$\begin{array}{rlccc} & & S_1 & S_2 & S_3 \\ i = 1. & S_1S_2S_3 & 1 & 1 & 1 \\ 2. & S_1S_2\overline{S}_3 & 1 & 1 & 0 \\ 3. & S_1\overline{S}_2S_3 & 1 & 0 & 1 \\ 4. & S_1\overline{S}_2\overline{S}_3 & 1 & 0 & 0 \\ 5. & \overline{S}_1S_2S_3 & 0 & 1 & 1 \\ 6. & \overline{S}_1S_2\overline{S}_3 & 0 & 1 & 0 \\ 7. & \overline{S}_1\overline{S}_2S_3 & 0 & 0 & 1 \\ 8. & \overline{S}_1\overline{S}_2\overline{S}_3 & 0 & 0 & 0 \end{array}$$

If we set $S_1 \leftrightarrow P,\ S_2 \leftrightarrow P \supset Q,\ S_3 \leftrightarrow Q$ and 'algebraically' eliminate P and Q we obtain the analogue of Boole's $D = 0$ equation, namely

$$S_1S_2\overline{S}_3 \vee S_1\overline{S}_2S_3 \vee \overline{S}_1\overline{S}_2S_3 \vee \overline{S}_1\overline{S}_2\overline{S}_3 \leftrightarrow f,$$

representing those alternands on $S_1,\ S_2,\ S_3$, which are impossible (by *logic*) on the data, hence contributing no probability. Striking out the corresponding rows, i.e., rows 2, 3, 7 and 8, produces

7. A basic tenet of this monograph. In §4.7 below we discuss the relationship between the customary set-theoretic and our logic-theoretic approach to probability.

	S_1	S_2	S_3
$S_1S_2S_3$	1	1	1
$S_1\overline{S}_2\overline{S}_3$	1	0	0
$\overline{S}_1S_2S_3$	0	1	1
$\overline{S}_1S_2\overline{S}_3$	0	1	0

which is the transpose of (5).

Since Nilsson's development is based on an analysis such as is illustrated in (3), and since we have seen that one can arrive at it by ordinary logical means, here too we believe the notion of 'worlds' is not essential. In fact, they disappear for Nilsson when he identifies (p. 73) sets of possible worlds with sets of truth values for sentences. We suggest going one step further and replace the sets of truth values by constituents (on the S_1, S_2, S_3), which are sentences.

We turn to the other items singled out from our initial quotation. With regard to item (i) the description of his 'probabilistic logic' is quite brief (*Nilsson 1986*, 73) :

> ... Since we typically do not know the ordinary (*true/false*) truth value of S in the actual world, it is convenient to imagine a logic that has truth values intermediate between *true* and *false* and, in this logic, define the truth value of S to be the probability of S. In the context of discussing uncertain beliefs, we use the phrases *the probability of S* and *the (probabilistic logic) truth value of S* interchangeably.

However the details of the 'logic' which Nilsson says it is convenient to imagine are not spelled out. All we are told is that it has ('truth') values between *true* and *false* and that the truth value of a sentence S is defined to be the probability of S. We have the following comments:

1. A (syntax) language for the logic is not explicitly stated. Nilsson implies (see our opening quotation from his paper) that the sentences are those of ordinary first-order logic.
2. There is no full account of the semantics and its relationship to the syntax. How does the value (probability) of a compound sentence relate to those of its components? And what if the sentence is a quantified expression?
3. There is no explicit definition of logical consequence. However Nilsson does introduce a procedure he calls "probabilistic entailment" which concerns "determining the probability of an arbitrary sentence given a set $\mathcal{B}$ of sentences and their probabilities".

If we restrict Nilsson's "logical systems" to sentential logic then his "determining the probability of an arbitrary sentence given a set $\mathcal{B}$ of sentences and their probabilities" is exactly Boole's General Probability Problem (see §§2.4, 2.5 above). Apparently when Nilsson wrote his paper he was unaware of any previous history, as his only relevant reference is to *Adams-Levine 1975* which, likewise, was unaware of earlier work. It is of interest to look at the problem from Nilsson's perspective, as we now do.

Suppose we are given a set $\mathcal{S}$ of sentences and a probability distribution over the sets of 'possible worlds' for $\mathcal{S}$ (i.e., a probability distribution over the consistent basic conjunctions of elements of $\mathcal{S}$). The combined relationship that obtains between the probabilities that accrue to the sentences of $\mathcal{S}$ by virtue of their syntactic structure and the probabilities of the distribution, is expressed by Nilsson as a matrix equation

$$\Pi = \mathbf{VP}, \tag{6}$$

where Π is a column matrix whose entries are the probabilities of the members of $\mathcal{S}$, where $\mathbf{V}$ is the matrix whose columns are the sets of truth values (in the form of 1's and 0's) which the sentences have in the possible worlds, and where $\mathbf{P}$ is the column matrix of the probability distribution values. Thus for the example $\mathcal{S} = \{P,\ P \supset Q,\ Q\}$ equation (6) is

$$\begin{bmatrix} p(P) \\ p(P \supset Q) \\ p(Q) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} \tag{7}$$

where the matrix $\mathbf{V}$ of 1's and 0's is the incidence matrix (5). To (6) and (7) one needs to add the conditions $p_i \geq 0$ and $\sum_i p_i = 1$. (This latter condition is replaceable by the addition of a logically true sentence to $\mathcal{S}$ which would then add a row of 1's to $\mathbf{V}$ and also a 1 to Π.) In the example (7) the first equation,

$$p(P) = p_1 + p_2 + 0p_3 + 0p_4,$$

'says' that the probability $p(P)$ is composed of the p_1-th part contributed by $S_1S_2S_3$ and the p_2-th part contributed fy $S_1\overline{S}_2\overline{S}_3$, and similarly for the other three equations. An analysis such as this first occurs in Boole *1854*, chapter XIX (modernized in *Hailperin 1965*).

The procedure which Nilsson employs to determine the probabilistic entailment of sentence S by a set of sentences $\mathcal{B}$ ('beliefs') with given probabilities, is to adjoin S to the set $\mathcal{B}$ and then for this enlarged set determine the region of possible probability values for S in the Π-space, subject to the contraint $\Pi = \mathbf{VP}$ (for the enlarged $\mathcal{B}$). Examination of this region gives Nilsson the allowable probability values for the sentence S. Thus, for the example $S = Q, \mathcal{B} = \{P, P \supset Q\}$, the matrix equation is (7). Because the

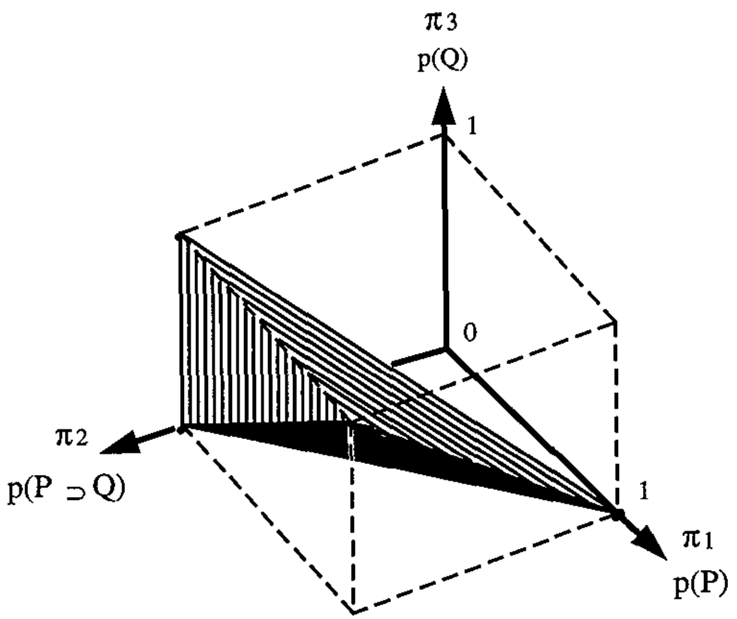

FIG. 3.71. The convex region of consistent probability values for P, $P \supset Q$, and Q.

small number of sentences involved Nilsson can view the problem geometrically. The region specified by the matrix equation is depicted in Figure 3.71 (taken from Nilsson, p. 76).

The solid has two non-vertical faces and two vertical faces (one hidden from view). From the equations of the non-vertical planes one readily deduces that if $(p(P),\ p(P \supset Q),\ p(Q))$ is a point of the solid then

$$p(P \supset Q) + p(P) - 1 \ \leq\ p(Q) \ \leq\ p(P \supset Q).$$

This same result, as an example of logical consequence in probability logic (generalized *modus ponens*), appeared in *Hailperin 1984*, 209–210, and is Theorem 4.51 below.

Nilsson observes that inconsistent probabilities could be assigned to premises (elements of $\mathcal{B}$). Any vertical line in the figure not intersecting the solid would intersect the $\pi_1\pi_2$ plane in a point $(p(P),\ p(P \supset Q))$ which furnishes no value for $p(Q)$. He also points out that the probability $p(Q)$ is not determined uniquely but only within bounds. He does not consider the question as to whether these bounds are best possible, i.e., give the strongest entailment result.

Our next comment concerns the third item of note listed by us at the beginning of this section. Nilsson states that the "semantical generalization of ordinary first-order logic" applies to any logical system for which the consistency of a finite set of sentences can be established. We take it that "established" means by an effective procedure. Since consistency for arbitrary finite sets of sentences of first-order logic is not an effective notion some restriction, e.g., to sentential logic or to monadic predicate logic, is needed; and, in the latter case, as we shall now see, no more than sentential logic is used.

The example of Nilsson's we wish to discuss is referred to as a "simple problem in first-order logic". It concerns the probabilistic entailment of $\exists z Q(z)$ by $\exists y P(y)$ and $\forall x(P(x) \supset Q(x))$. Using Skolem functions he finds the consistent assignments for the three sentences to be (using our form)

	$S_1S_2S_3$	$S_1\overline{S}_2S_3$	$S_1\overline{S}_2\overline{S}_3$	$\overline{S}_1S_2S_3$	$\overline{S}_1S_2\overline{S}_3$
$S_1(\exists y P(y))$	1	1	1	0	0
$S_2(\forall x(P(x) \supset Q(x)))$	1	0	0	1	1
$S_3(\exists z Q(z))$	1	1	0	1	0

from which he derives

$$p(\exists y P(y)) + p(\forall x(P(x) \supset Q(x)) - 1 \ \leq\ p(\exists z Q(z)) \ \leq\ 1. \tag{8}$$

However the problem is readily expressible on the sentential level. We rewrite the sentences (using obvious abbreviations) as follows:

$$\begin{aligned}
\exists y P(y) &= \exists x(P(x)Q(x) \vee P(x)\overline{Q(x)}) \\
&= \exists(PQ) \vee \exists(P\overline{Q}) = A \vee B \\
\forall x(P(x) \supset Q(x)) &= \neg\exists x(P(x)\overline{Q(x)}) \\
&= \neg\exists(P\overline{Q}) = \neg B \\
\exists z Q(z) &= \exists x(P(x)Q(x) \vee \overline{P(x)}Q(x)) \\
&= \exists(PQ) \vee \exists(\overline{P}Q) = A \vee C.
\end{aligned}$$

The set is then of the form

$$\mathcal{S} = \{A \vee B,\ \neg B,\ A \vee C\}$$

and produces the result (8) without reference to the quantifier structure of A, B, or C.

What we have shown here in this particular case is true in general for monadic predicate sentences. For such sentences consistency can be determined in a domain of individuals having no more that 2^n individuals, where n is the number of distinct predicates present. (See, e.g., *Hilbert-Bernays 1968*, 194.) Thus all quantification sentences can be replaced by finite conjunctions or alternations, which then converts the problem to the sentential (i.e., non-first-order) level. (Systematic methods for carrying out the reduction—as in our working of Nilsson's example—can be gleaned from *Quine 1972*, §§18–24.)

Although realizing that the problem of probabilistic entailment can be treated by linear programming methods, Nilsson nevertheless doesn't make use of the results and methods of this subject. For example, he remarks (p. 86) that his method can be extended to include ("belief") sentences with given upper and lower bound probabilities and not just exact values. But rather than including these as inequality constraints (as in our *1965*, §6) he suggests instead calculating bounds for the entailed S by first using one set of extreme values, and then again for the other set of extreme values.

There is a brief discussion of conditional probabilities in probabilistic logic. The method of treatment suggested is to write the conditional probability as a quotient of unconditional probabilities and to find the bounds for each separately. We contend this cannot give best results in all cases. See, for example, the criticism of a similar procedure of Boole's in our *1986*, 371–372; it contains in its §6.7 a quite different method, which doesn't separate the numerator and denominator probabilities but treats the quotient as a whole. Our chapter 5 below has a detailed treatment of the method.

A number of papers based on Nilsson's subsequently appeared in various journals devoted to operations reserach, artificial intelligence, expert systems and the like. We cite only *Georgakopoulos, Kavvadias and Papadimitriou 1988* and *Jaumard, Hansen and Poggi de Aragão 1991*, these being specially concerned with the large scale computations that arise when there are many variables and conditions. The latter of these two papers includes in its bibliography references to other papers that were spawned by Nilsson's.

§3.8. Probability logic of Scott and Krauss (*1966*)

The paper we discuss in this section antedated both that of Adams-Levine (§3.6) and that of Nilsson (§3.7). According to the authors it was inspired by *Gaifman 1964*, which introduced and investigated probability measures on (finitary) first-order languages. Scott and Krauss frame their results for *infinitary* first-order languages, i.e., languages which, among

other things, allow concatenation of transfinitely many (but less than ω_1) symbols to form a formula. We will not engage in this much generality. With appropriate modifications the Scott-Krauss results which are of interest to us can be specialized to finitary (i.e., ordinary) first-order or to sentential languages. Restricting attention to the sentential case, as we will now be doing, while excising the substantive results of their paper, will enable us to bring out a difference with what we present in the next chapter.

A *probability model* (for the sentential case) is an $(n+2)$-tuple of the form

$$\langle A_1^0, \ldots, A_n^0, \mathcal{A}, m\rangle$$

where

(i) $A_1^0, \ldots, A_n^0$ are sentences (of an interpreted language),
(ii) $\mathcal{A}$ is a Boolean algebra (or, also, the set of its elements),
(iii) m is a probability (measure) on $\mathcal{A}$ which is strictly positive ($m(a) = 0$ if and only if a is the $\mathbf{0}$ of the Boolean algebra), and
(iv) there is a mapping of the sentences A_i^0 into $\mathcal{A}$, the $\mathcal{A}$-value assigned to A_i^0 being denoted by 'a_i'.

(Here $\mathcal{A}$ generalizes the role played by the two-element Boolean algebra in the case of verity logic.)

Let $\mathcal{S}$ be a formal sentential language with sentential variables $A_1, \ldots, A_n, \ldots$ and connectives $\neg$, $\vee$, $\wedge$; let $\mathcal{S}_n$ be the part of $\mathcal{S}$ all of whose formulas have variables contained in the set $\{A_1, \ldots, A_n\}$. Analogous to 'ϕ holds in a model $\mathfrak{A}$' for verity logic, there is the following definition of 'ϕ holds in a probability model $\mathcal{U}$ with probability α'.

First, for a probability model $\mathcal{U} = \langle A_1^0, \ldots, A_n^0, \mathcal{A}, m\rangle$ one associates a value in $\mathcal{A}$ for each $\phi \in \mathcal{S}_n$ when the $A_1, \ldots, A_n$ are thought of, or interpreted as, $A_1^0, \ldots, A_n^0$. This is accomplished by extending the assignment (iv) via a recursively defined *valuation* function h:

$$\begin{aligned} h(A_i) &= a_i, \quad \text{for } i = 1, \ldots, n \\ h(\neg\phi) &= \mathbf{1} \sim h(\phi) \text{ (i.e., the complement of } h(\phi)), \\ h(\phi \wedge \psi) &= h(\phi) \cap h(\psi), \\ h(\phi \vee \psi) &= h(\phi) \cup h(\psi). \end{aligned}$$

It can be shown that if ϕ is logically valid ($\vdash \phi$) then $h(\phi) = \mathbf{1}$ ($\mathbf{1}$ is the Boolean unit), and if ϕ and ψ are logically equivalent ($\vdash \phi \leftrightarrow \psi$) then $h(\phi) = h(\psi)$. Setting $\mu_{\mathcal{U}}(\phi) = m(h(\phi))$ defines what Gaifman (*1964*, 2) refers to as a probability measure (on $\mathcal{S}_n$). Then

$$\mu_{\mathcal{U}}(\phi) = \alpha$$

expresses

ϕ holds in the probability model $\mathcal{U}$ with probability α.

Thus h maps sentences into the Boolean algebra $\mathcal{A}$, whose measure m then endows sentences of $\mathcal{S}_n$ with a (probability) value. In this simplified situation the Boolean algebra $\mathcal{A}$ in the definition of probability model serves no particular function except as a carrier of the probability measure. In its place one could just as well use the Boolean algebra of sets of formulas defined for $\mathcal{S}_n$; the elements of this Boolean algebra are the equivalence classes of formulas of $\mathcal{S}_n$, modulo logical equivalence.

Preparatory to introducing the definition of probability (logical) consequence Scott and Krauss define a probability assertion, i.e., an assertion of probability logic. The notion requires that the language include symbols for the (elementary) algebra of real numbers (= the theory of real-closed fields); these are: a binary relation symbol $\leq$, binary function symbols $+$, $\times$, and the individual constants 0, +1, −1. These symbols are interpreted over the real numbers in the usual manner.[8] Note that in this language the only real numbers that can be explicitly symbolized are the integers. An *algebraic formula* is a quantifier-free formula of the language. It can be shown (*Tarski 1948*, 18) that any algebraic formula is equivalent in real algebra to an alternation of conjunctions of polynomial inequations of the form $p \geq 0$ or $p > 0$, i.e., to a positive logical function of such inequations. A *probability assertion* then is an $(N+1)$-tuple $\langle \Phi, \phi_1, \ldots, \phi_N \rangle$, where Φ is an algebraic formula with N free variables and $\phi_1, \ldots \phi_N \in \mathcal{S}_n$. The definition of a probability model $\mathcal{U}$, in which for each formula ϕ of $\mathcal{S}_n$ there is an associated value $\mu_{\mathcal{U}}(\phi)$ is now to be extended to probability assertions: a probability model $\mathcal{U}$ is a *probability model of (an assertion)* $\langle \Phi, \phi_1, \ldots, \phi_N \rangle$ if the N-tuple $\langle \mu_{\mathcal{U}}(\phi_1), \ldots, \mu_{\mathcal{U}}(\phi_N) \rangle$ satisfies Φ in the reals, i.e., that Φ is true of $\langle \mu_{\mathcal{U}}(\phi_1), \ldots, \mu_{\mathcal{U}}(\phi_N) \rangle$. Now for the Scott-Krauss definition of consequence.

> If Σ is a set of probability assertions and Ψ is a probability assertion, then Ψ is a *probability consequence of* Σ iff every probability model of all assertions in Σ is also a probability model of Ψ. Ψ is a *probability law* of $\mathcal{L}$ if Ψ is a probability consequence of the empty set of assertions.

Scott and Krauss establish the following significant result for their rich (infinitary) language $\mathcal{L}$ (*1966*, 243):

8. We trust that the reader will not be confused by our using the same notation for the symbols of the theory of real-closed fields as for their interpretation in the reals.

Theorem 6.7. *Let $\langle \Phi, \phi_0, \ldots, \phi_{N-1} \rangle$ be a probability assertion of $\mathcal{L}$ such that the free variables of Φ are $\lambda_0, \ldots, \lambda_{N-1} \rangle$; further $\vdash \neg(\phi_i \wedge \phi_j)$ if $i \neq j$, and $\vdash \bigvee_{i<N} \phi_i$. Let $I = \{i < N : \vdash \neg\phi_i\}$. Then $\langle \Phi, \phi_0, \ldots, \phi_{N-1} \rangle$ is a probability law of $\mathcal{L}$ iff the sentence*

$$\forall \lambda_0 \ldots \forall \lambda_{N-1} \left[\bigwedge_{i \in I} \lambda_i = 0 \wedge \bigwedge_{i<N} \lambda_i \geq 0 \wedge \lambda_0 + \cdots + \Lambda_{N-1} = 1 \rightarrow \Phi \right]$$

is a theorem of real algebra.

In other words, if it is provable that $\phi_0 \vee \cdots \vee \phi_{N-1}$ and that the ϕ_i are pairwise disjoint, then the probability assertion

$$\langle \Phi, \phi_0, \ldots, \phi_{N-1} \rangle$$

is a probability law if and only if Φ is true of all sets of non-negative real numbers $\lambda_0, \ldots, \lambda_{N-1}$ which sum to 1, with the λ_i corresponding to those ϕ_i's provably false being set equal to 0.

For a sentential language $\mathcal{S}_n$ the result takes on a simpler form. Any formula ϕ of $\mathcal{S}_n$, if not logically false, is equivalent to a logical sum of the form $\bigvee K_j$, the K_j being constituents on $A_1, \ldots, A_n$. Since constituents are mutually exclusive we have, for any probability measure μ,

$$\mu(\phi) = \mu(\bigvee K_j) = \sum \mu(K_j).$$

If $\Phi(\mu(\phi_1), \ldots, \mu(\phi_N))$ is an algebraic formula with arguments $\mu(\phi_i)$ $(i = 1, \ldots, N)$ then on replacing each $\mu(\phi_i)$ by its equal sum $\sum_{\phi_i} \mu(K_j)$ (or 0, if ϕ_i is logically false) produces (after some simple algebra) an algebraic formula of the form $\Psi(\mu(K_1), \ldots, \mu(K_{2^n}))$. Note that when the ϕ_i of the Scott-Krauss Theorem 6.7 are the constituents $K_1, \ldots, K_{2^n}$ then the hypotheses are satisfied and also that the set I is empty. Thus we have for $\mathcal{S}_n$ the following corollary of Theorem 6.7:

Corollary. *Let $\langle \Phi, K_1, \ldots, K_{2^n} \rangle$ be a probability assertion of $\mathcal{S}_n$ such that the free variables of Ψ are $k_1, \ldots, k_{2^n}$ and $K_1, \ldots, K_{2^n}$ are the constituents on $A_1, \ldots, A_n$. Then $\langle \Psi, K_1, \ldots, K_{2^n} \rangle$ is a probability law of $\mathcal{S}_n$ if and only if the sentence*

$$\forall k_1 \ldots \forall k_{2^n} \left[\bigwedge_{1 \leq j \leq 2^n} (k_j \geq 0) \wedge k_1 + \cdots + k_{2^n} = 1 \rightarrow \Psi \right]$$

is a theorem of real algebra.

As a consequence, for sentential probability logic one can use a simpler definition of a probability model, namely a model is an assignment to $K_1, \ldots, K_{2^n}$ of any set of non-negative (or positive, if strict additivity is desired) real numbers $k_1, \ldots, k_{2^n}$ which sum to 1. This is the definition used in our *1984*—though with a different definition of logical consequence—and in §4.3 below. We shall be comparing the two definitions in §4.4.

Chapter 4

Formal Developments

§4.1. What is *a* logic?

Considering the conventional nature of words and that subject matter changes over the course of time, this question is a possibly impossible one to answer. A brief examination of a piece of literature lying at hand produced the following *score* of systems referred to as 'logic': alternative, Aristotelian, classical, combinatory, deontic, epistemic, erotetic, free, fuzzy, inductive, infinitary, intuitionistic, many-valued, modal, non-Aristotelian, probability, quantum, relevance, strict implication and temporal—not to mention variants and subdivisions. Rather than attempting a definitive characterization we shall describe, informally, general features which we think something should have in order to be called a logic.

There should be a formal language and an associated semantics. The internal structure of the atomic sentences of the language may or may not be specified, but the logical syntax of the language should be completely specified. This requirement involves the notion of a logical constant. In this book, aside from peripheral matters, we shall be dealing only with sentential (propositional) languages. For such languages, the internal structure of the atomic sentences playing no role, the only logical constants are the sentential connectives. The logical properties of connectives are determined by the associated semantics, which specifies 'truth'-values of compounds in terms of the 'truth'-values of their atomic sentences. (In keeping with our general viewpoint we shall now use the term *semantic value* in place of 'truth'-value.) It has generally been assumed that the semantic value of a compound sentence has to be uniquely determined by the semantic values of its atomic sentences. This feature will not be present in our probability logic.

Up until almost the middle of this century the customary form in which a