unproven fact stated on p. 423

GLENN SHAFER

A THEORY OF STATISTICAL EVIDENCE

TABLE OF CONTENTS

Harper and Hooker (eds.), Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II, 365–436.

1. Evidence

There are at least two ways in which the impact of evidence on a proposition may vary. On the one hand, there are various possible degrees to which the evidence may support the proposition: taken as a whole, it may support it strongly, just a little, or not at all. On the other hand, there are various possible degrees to which the evidence may cast doubt on the proposition: taken as a whole, it may cast serious doubt on it, thus rendering it extremely doubtful or implausible; it may cast only moderate doubt on it, thus leaving it moderately plausible; or it may cast hardly any doubt on it, thus leaving it entirely plausible.

1.1. Degrees of support and plausibility

In this essay, I formally distinguish these two aspects of the evidence's impact. I say that the evidence supports a proposition to a certain extent, thus endowing it with a certain *degree of support*, and that it casts doubt on it to a certain extent, thus endowing it with a certain *degree of plausibility*.

I approach these degrees of support and plausibility with two ambitions. First, I hope that in some situations they can actually be represented by numbers. And secondly, I hope that in these situations such numerical degrees of support and plausibility for relevant propositions will be sufficient to *completely summarize* the evidence's impact on our knowledge and opinion.

A proposition's degree of support and its degree of plausibility are obviously related, and it might seem that they are so strongly related that the one should determine the other. But they are not. The fact is that while a high degree of support does imply a high degree of plausibility, a low degree of support is compatible both with a low degree of plausibility and with a high degree of plausibility. If the evidence supports the proposition not at all and casts a great deal of doubt on it, then it endows it with a low degree of support and a low degree of plausibility. But if the evidence fails to provide much support for the proposition and also fails to cast much doubt on it, then it endows it with a low degree of support and yet leaves it with a high degree of plausibility. Actually, this latter situation is all too common, for it arises whenever the evidence is scanty. When there is little evidence bearing on a proposition, that proposition

cannot be said to be supported by the evidence, but it is plausible even in light of the evidence.

I want these degrees of support and plausibility to be numbers. What numbers?

Consider first the range of numbers that we might want for degrees of support.

At one extreme, when there is no support for a proposition, we will want to say that its degree of support is zero. At the other extreme, we will want a maximum degree of support, corresponding to the case where the evidence establishes the proposition for certain. A convenient convention is to set this maximum degree of support equal to one. So when we measure the degree of support for a proposition we will assign the proposition a number between zero and one.

This same scale from zero to one also seems appropriate for degrees of plausibility.

A proposition will have degree of plausibility zero when the evidence is conclusively against it, and degree of plausibility one when there is no evidence against it.

1.1.1. *The Support and Plausibility Functions*

We are usually interested in degrees of support and plausibility for more than one proposition at a time. For example, when we are concerned with the true value of some quantity θ, we are interested in any proposition that asserts that the true value is included in a given subset of the set of possible values.

Denoting the set of possible values by Θ, the propositions of interest are precisely those of the form 'The true value of θ is in A', where A is a subset of Θ.

Hence the propositions of interest are in a one-to-one correspondence with the subsets of Θ, and for the sake for convenience we can 'identify' them with these subsets.

So denoting the set of all subsets of Θ, or the *power set* of Θ, by the symbol 2^Θ, we can describe our problem as that of specifying two functions on 2^Θ. First we want a function

$$S: 2^\Theta \to [0, 1]$$

such that $S(A)$ is the degree of support for the subset A. And secondly, we

want a function

$$PL: 2^{\Theta} \to [0, 1]$$

such that $PL(A)$ is the degree of plausibility of A.

This formalism may seem fairly special. For we are often interested in propositions that do not deal with the value of a numerical quantity. But it becomes quite general if we allow θ to be a 'parameter' that takes possibly non-numerical values. For example, we might let θ be 'the date and place of origin of the relic in my hand'. In this case, the possible values of θ would be pairs, each pair consisting of a date and a place.

Whatever θ is, it should be noted that whenever a subset $A \subset \Theta$ is thought of as a proposition, its complement $\bar{A}$, the set of all elements of Θ not in A, must be thought of as the negation of that proposition. Notice also that the empty set $\emptyset$ is in 2^{Θ}; it corresponds to the proposition that is necessarily false, for the true value of θ cannot be in $\emptyset$. And the entire set Θ is also in 2^{Θ}; it corresponds to the proposition that is necessarily true, for by assumption the true value of θ is in Θ.

1.1.2. *The Relation Between S and PL*

I have already pointed out the relation between a proposition's degree of support and its degree of plausibility: high support implies high plausibility, but low support is compatible with both low and high plausibility. We can formalize this relation by requiring that the degree of plausibility be at least as great, but possibly greater than the degree of support. In symbols:

$$(1) \qquad S(A) \leqslant PL(A)$$

for each $A \in 2^{\Theta}$. In words: plausibility is easier to come by than support.

A fundamental relation between support and plausibility can be discerned when one compares the degree of plausibility of a proposition $A \in 2^{\Theta}$ with the degree of support for its negation $\bar{A}$. Recall that a proposition A is plausible to the extent that the evidence fails to cast doubt on it. But casting doubt on A is really the same thing as supporting $\bar{A}$. Hence A is plausible to the extent that $\bar{A}$ fails to be supported; $PL(A)$ is large to the extent that $S(\bar{A})$ is small.

Since both $PL(A)$ and $S(\bar{A})$ are measured on a scale from zero to one, the most natural way to make this relation precise is to set

$$(2) \qquad PL(A) = 1 - S(\bar{A})$$

for all $A \in 2^{\Theta}$.

This relation implies not only that we can obtain the function PL from knowledge of S, but also that we can obtain S from knowledge of PL. For (2) implies that

$$(3) \qquad S(A) = 1 - PL(\bar{A})$$

for all $A \in 2^{\Theta}$. Hence the functions PL and S convey exactly the same information.

1.1.3. *Elementary Rules for S and PL*

It is worth noting that the relations (1) and (2) imply that

$$S(A) + S(\bar{A}) \leqslant 1$$

for all $A \in 2^{\Theta}$. Verbally: it is impossible for both a proposition and its negation to be well supported. Similarly, (1) and (3) imply that

$$PL(A) + PL(\bar{A}) \geqslant 1$$

for all $A \in 2^{\Theta}$. Verbally: for every proposition, either it, its negation or both must be fairly plausible.

There are several other rules that S and PL should obey. For one thing, the elements $\emptyset$ and Θ of 2^{Θ} are rather special. No matter what the evidence is, $\emptyset$ is impossible and hence $S(\emptyset) = PL(\emptyset) = 0$. Similarly, Θ is always taken to be certain, and even in the absence of any evidence we would set $S(\Theta) = PL(\Theta) = 1$. The function S ought also to obey the rule of monotonicity:

$$\text{If} \quad A \subset B, \quad \text{then} \quad S(A) \leqslant S(B).$$

This rule is unavoidable, for when $A \subset B$, any support for the value of θ being in A is also support for the value of θ being in B. Finally, the rule of monotonicity for S implies exactly the same rule for PL:

$$\text{If} \quad A \subset B, \quad \text{then} \quad PL(A) \leqslant PL(B).$$

The rules for S and PL are summarized in Table I.

TABLE I

Rules for S and PL. To the right of each rule for PL is the corresponding rule for S, based on the relation $S(A)=1-PL(\bar{A})$

Rules for plausibility	Rules for support
$PL(\emptyset)=0$	$S(\Theta)=1$
$PL(\Theta)=1$	$S(\emptyset)=0$
If $A\subset B$, then $PL(A)\leq PL(B)$	If $A\subset B$, then $S(A)\leq S(B)$
$PL(A)+PL(\bar{A})\geq 1$	$S(A)+S(\bar{A})\leq 1$

1.2. The case of two alternatives

The simplest support and plausibility functions occur when Θ consists of only two alternatives; say $\Theta=\{\theta_1, \theta_2\}$. In this case the support function $S: 2^\Theta \to [0, 1]$ will be completely determined by two numbers: $S(\{\theta_1\})=s_1$ and $S(\{\theta_2\})=s_2$. And since $\{\theta_2\}=\overline{\{\theta_1\}}$, these two numbers must obey $s_1+s_2\leq 1$.

TABLE II

The general form for S and PL when $\Theta=\{\theta_1, \theta_2\}$ $(s_1+s_2\leq 1)$

A	$S(A)$	$PL(A)$
$\emptyset$	0	0
$\{\theta_1\}$	s_1	$1-s_2$
$\{\theta_2\}$	s_2	$1-s_1$
Θ	1	1

1.2.1. *No Evidence*

It is generally difficult if not impossible to actually assess the evidence and arrive at numerical degrees of support and plausibility. But in one case it is easy – the case where there is no evidence. When there is no evidence none of the propositions in 2^Θ can be supported, and hence all must have degree of support zero. Dually, none can have doubt cast on them and hence all must have degree of plausibility one. Of course, we must make two exceptions: $\emptyset$ being logically impossible, we must have $PL(\emptyset)=0$; and Θ being logically certain, we must have $S(\Theta)=1$. So when there is no evidence we obtain the *vacuous* support function, which assigns every proposition except Θ support zero, and the *vacuous* plausibility function, which assigns every proposition except $\emptyset$ plausibility one.

In terms of Table II, such a vacuous support function would be represented by setting $s_1=s_2=0$. Let me give a concrete example. My friend the art collector shows me a vase and tells me that it has been represented as a product of the Ming dynasty. He asks me what I think – is it genuine or is it counterfeit? Now the only thing I know about vases is that flowers can be kept in them, and the only thing I know about the Mings is that they were not Frenchmen. Surely you will agree that I have no evidence and hence should adopt the vacuous support function, shown in Table III.

TABLE III

The vacuous support function when $\Theta=\{\text{genuine, counterfeit}\}$

A	$S(A)$	$PL(A)$
$\emptyset$	0	0
{genuine}	0	1
{counterfeit}	0	1
Θ	1	1

1.2.2. *Conflicting Evidence*

The case of no evidence is easy. But we usually have some evidence. Indeed, there is often so much evidence that we can find some against each possibility.

Consider again the question of whether the vase is genuine or counterfeit: $\Theta=\{\text{genuine, counterfeit}\}$. By definition it is one or the other, but I will not be surprised if I find some evidence pointing to its being genuine and some evidence pointing to its being counterfeit. Suppose I do. For the sake of concreteness, suppose my evidence comes from a study of the painting on the vase. My evidence for the vase's being genuine may be the particular skill exhibited in the intricate design, a skill that is not believed to have survived the Ming period. And my evidence for its being counterfeit may be the presence of a certain pigment previously believed to have been unavailable during the Ming period. Suppose both of these items of evidence seem to be fairly weighty, but equally weighty. Then the total body of evidence can be aptly described as internally conflicting: there may be quite a bit of it, but is points in both directions at once.

What does the plausibility function look like? Both alternatives have evidence against them and thus doubt cast on them; hence neither remains completely plausible. Let us suppose they both retain plausibility 3/4. Then we obtain the support and plausibility functions shown in Table IV. Notice that the set $\Theta = \{\text{genuine, counterfeit}\}$ retains plausibility one even though neither of its individual elements are that plausible any more.

Table IV might strike you as a complicated way of saying nothing. "Sure you have all that evidence", you might argue. "But it cancels itself

TABLE IV

A support function that indicates internally conflicting evidence

A	$S(A)$	$PL(A)$
$\emptyset$	0	0
{genuine}	1/4	3/4
{counterfeit}	1/4	3/4
Θ	1	1

out. Why should you say that the evidence supports both alternatives to a degree of 1/4? It would be simpler to say that the tendencies to provide support in the two opposite directions cancel each other completely, leaving support zero and plausibility one for both alternatives". In other words, you might argue that this precisely balanced conflicting evidence really comes down to the same thing as no evidence at all.

Admittedly, not all the vocabulary we use nowadays to discuss evidence is adapted to distinguishing between a lack of evidence and the presence of conflicting evidence. For example, conflicting evidence does no better than no evidence in providing us with 'information' or helping us make a 'decision'. But the difference between no evidence and conflicting evidence is both real and practical and should be basic to any theory of evidence.

1.2.3. *The Combination of Evidence*

The importance of the difference between no evidence and conflicting evidence emerges clearly when we undertake to combine the evidence we already have with new evidence. Suppose, for example, that we have the conflicting evidence summarized in Table IV, and that we then obtain

new evidence strongly in favor of the vase's being genuine. Indeed, suppose the new evidence supports that alternative fairly strongly and does not cast any doubt on it at all. Then this new evidence, taken by itself or combined with 'no evidence', will result in $S(\{\text{genuine}\})$ being fairly high and in $PL(\{\text{genuine}\})$ being equal to one. But when this new evidence is combined with the previous conflicting evidence, the result will be different. Certainly the new evidence will shift the balance in favor of the vase's being genuine, and perhaps it will raise the degree of support for that alternative. But the original evidence against the vase's being genuine (the suspicious pigment) will remain part of the combined evidence, and hence $PL(\{\text{genuine}\})$ will not be equal to one.

Let me make the example more concrete. Say the new evidence is the testimony of an expert who has analyzed the chemical composition of the clay in the vase and has concluded that it could only have come from the Ming period. Now we may put some faith in his expertise and his honesty, so his testimony will provide positive support for the vase's being genuine. But experts can be wrong, so we will not regard his testimony as conclusive evidence. Suppose we think it provides a degree of support of 1/2 for the vase's being genuine. Then when considered alone or combined with no evidence it will produce the support and plausibility functions given in Table V.

TABLE V

The support function based on the new evidence

A	$S(A)$	$PL(A)$
$\emptyset$	0	0
{genuine}	1/2	1
{counterfeit}	0	1/2
Θ	1	1

Let us consider the degrees of support that might be expected to result from the combination of the old evidence represented by Table IV with the new evidence represented by Table V. It seems reasonable that the combination of the two bodies of evidence should produce a fairly high degree of support for the vase's being genuine, perhaps higher than the

degree of support in either table. On the other hand, the degree of support for the vase's being counterfeit ought to fall between the 0 in Table V and the 1/4 in Table IV. There will be positive support, deriving from the old evidence, but the one-sidedness of the new evidence will erode it considerably. Numbers in Table VI seem to meet these general requirements.

TABLE VI

The support function resulting from combination

A	$S(A)$	$PL(A)$
$\emptyset$	0	0
genuine	4/7	6/7
counterfeit	1/7	3/7
Θ	1	1

1.2.4. *Lambert's Rule*

Actually, Table VI can be obtained from Tables IV and V by the application of a simple rule that was first proposed by J. H. Lambert in 1764 and was rediscovered by A. P. Dempster in 1966.

In order to describe the application of this rule, we need to learn how to represent a support function over two alternatives by a mass that is uniformly distributed over a line segment. This is done in Figure 1 for the support function in Table IV.

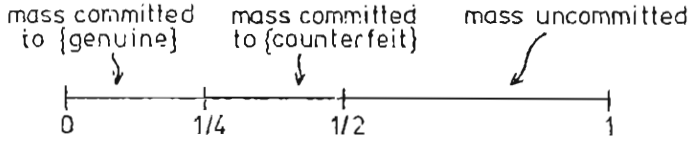

Fig. 1.

In that figure, 1/4 of the mass is committed to {genuine}, corresponding to the degree of support of 1/4 for that alternative, and 1/4 is committed to {counterfeit}, corresponding to the degree of support of 1/4 for that alternative. The support function in Table V is similarly represented in Figure 2. Since the degrees of support for two alternatives must always

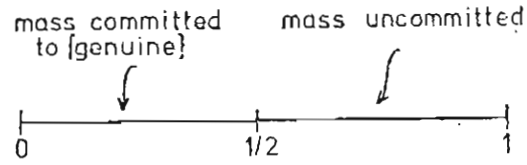

Fig. 2.

add to some number less than or equal to one, any support function over two alternatives can be represented in this way.

In order to combine these two support functions, we combine the two line segments orthogonally, obtaining the square shown in Figure 3. The division of the line segment of Figure 1 into three pieces then induces a division of the square into three vertical strips, while the division of the line segment of Figure 2 into two pieces induces a division of the square into two horizontal strips. Altogether the square is thus partitioned into $3 \times 2 = 6$ rectangles which I have labeled with the letters A through F.

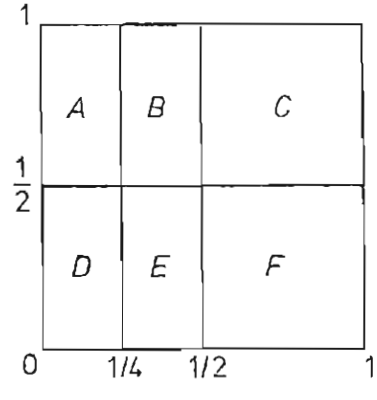

Fig. 3.

Let us consider how each of these six rectangles is affected by the two support functions. Rectangle A is committed to {genuine} by the first support function – a commitment that is not challenged by the second support function; hence it may be counted as committed to {genuine}. Rectangle B is similarly committed to {counterfeit}. Rectangle C, not being committed by either of the two, must be counted as uncommitted.

Rectangle D is committed to {genuine} by the first support function, and the second concurs. But for rectangle E there is a conflict: the first support function would commit it to {genuine} and the second would commit it to {counterfeit}; hence it cannot be counted at all. Finally, rectangle F is committed to {genuine} by the second support function, and this is not challenged by the first one (Figure 4).

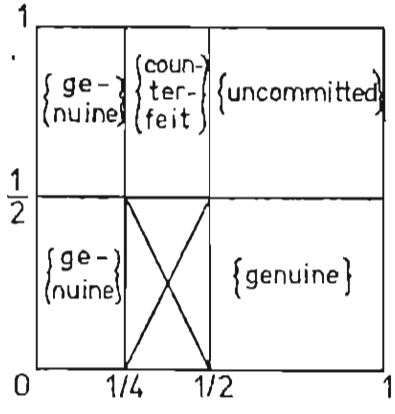

Fig. 4.

The net result, then, is that 4/8 of the square is committed to {genuine}, 1/8 is committed to {counterfeit}, 2/8 is uncommitted, and 1/8 cannot be counted. So of the 7/8 of the square that can be counted, 4/7 is committed to {genuine}, 1/7 is committed to {counterfeit}, and 2/7 is uncommitted. This result corresponds to the numbers in Table VI.

It should be obvious that the procedure used here can be used to combine any pair, or any larger number, of support functions over the same two alternatives. The rule is so simple as to seem almost silly, but its results are always intuitively reasonable.

1.2.5. *Caveat*

Lambert formulated his rule only for the case of two alternatives, but as we will see in §3, Dempster's more general rule of combination can be applied when Θ has any number of elements, provided that the support functions satisfy certain further conditions. This greater generality is important, for the use of the rule for only two alternatives has severe limitations.

The fact is that our evidence often refers to rather large sets of alternatives and loses much of its precision when we narrow our attention to a single dichotomy. In the case of the vase, for example, we might obtain new evidence not bearing directly on the vase's origin, but merely tending to impeach one of the sources of our old evidence. Such new evidence certainly ought to affect the degrees of support and plausibility based on our old evidence, but it can hardly do so by means of Lambert's rule. For it will produce only a vacuous support function when applied directly to the dichotomy $\Theta = \{\text{genuine, counterfeit}\}$. In order to combine the old and new evidence properly, we would have to apply Dempster's rule to support functions over a set Θ large enough to distinguish not only whether the vase is genuine but also whether the various sources of evidence are trustworthy.

In general, then, Dempster's rule ought always to be applied to support functions over sets Θ that are large enough to make all relevant distinctions. If Θ represents too coarse a division of the possibilities, then the result of the combination may be inaccurate. We can achieve our ambition of completely summarizing the evidence by a support function only if we make Θ sufficiently large.

1.3. Consonance and dissonance

When Θ consists of two alternatives, a support function that awards positive degrees of support to both alternatives betrays internal conflict or *dissonance* in the evidence. Dissonance can be revealed in a similar way when Θ is larger. Indeed, whenever $A \in 2^\Theta$ and both $S(A)$ and $S(\bar{A})$ are positive, the evidence and the support function S are clearly internally conflicting and hence dissonant. Is this the only way in which a support function can betray dissonance in the evidence?

It is not the only way. As we will see in the following example, a support function can betray dissonance even when it avoids the outright conflict involved in supporting both sides of a dichotomy.

1.3.1. *Another Vase*

Suppose θ is the exact place of origin of a Chinese vase. We may take Θ to be the set of all places in China, or we may abstract a bit and take Θ to

be the set of all points on a map of China; this will give us an infinite number of 'places' in China, many of them very close together. And suppose the evidence is dissonant.

What might it mean for the evidence to be dissonant in this example? Consider a subset A of the map and two subsets A_1 and A_2 of A such that $A_1 \cap A_2 = \emptyset$ and $A_1 \cup A_2 = A$; say A is the province of Honan and A_1 and A_2 are western and eastern Honan, respectively. For the evidence to be dissonant with respect to A_1 and A_2 would mean that some evidence points in one direction, say towards A_1 and away from A_2, while other evidence points in the other direction, towards A_2 and away from A_1. It is not hard to imagine how such conflicting evidence might arise; some aspects of the vase's design might resemble other pottery from sites in eastern Honan, while other aspects might seem more likely to have come from the west.

But we must draw some distinctions. Since Honan does not make up the whole map, evidence pointing towards western Honan is not exactly the same thing as evidence pointing away from eastern Honan. Hence, we can distinguish two different kinds of dissonance. First, there might be some evidence pointing towards, and hence supporting western Honan, and other evidence pointing towards, and hence supporting eastern Honan. Since eastern and western Honan are disjoint, this would certainly be a conflict. Secondly, there might merely be some evidence pointing away from western Honan without pointing away from eastern Honan and other evidence pointing away from eastern Honan without pointing away from western Honan. In other words, there might be some evidence casting doubt on western Honan but not on eastern Honan and other evidence casting doubt on eastern Honan but not on western Honan. In this case, the evidence taken as a whole would have to be considered dissonant, even though we might not want to say that it is in clear conflict with itself. Let us consider both kinds of dissonance in turn.

In the first case, we have positive support both for A_1 = western Honan and for A_2 = eastern Honan. These two subsets are disjoint: $A_1 \cap A_2 = \emptyset$. So dissonance is revealed in this case by the fact that

$$(4) \qquad A_1 \cap A_2 = \emptyset, \qquad S(A_1) > 0 \quad \text{and} \quad S(A_2) > 0.$$

Notice that the disjointness of A_1 and A_2 implies that $A_2 \subset \bar{A}_1$; hence

$S(\bar{A}_1) \geq S(A_2) > 0$. So whenever (4) occurs, we have

$$(5) \qquad S(A_1) > 0 \quad \text{and} \quad S(\bar{A}_1) > 0.$$

Hence (4) is merely another way of saying that there is positive support for both sides of a dichotomy.

Fig. 5.

The second kind of dissonance, which is weaker but more interesting, is best described in terms of the plausibilities $PL(A)$, $PL(A_1)$ and $PL(A_2)$. There may be some doubt cast on Honan as a whole and that doubt will also apply to eastern Honan and to western Honan, thus pushing down all three values $PL(A)$, $PL(A_1)$ and $PL(A_2)$. But in the case we are describing, there is evidence casting additional doubt on western Honan but not on eastern Honan and hence not on Honan as a whole, and also evidence casting additional doubt on eastern Honan, but not on western Honan and hence not on Honan as a whole. Hence, both $PL(A_1)$ and $PL(A_2)$ will be pushed down farther than $PL(A)$; we will have both $PL(A_1) < PL(A)$ and $PL(A_2) < PL(A)$. So the occurrence of the relations

$$(6) \qquad PL(A_1) < PL(A_1 \cup A_2) \quad \text{and} \quad PL(A_2) < PL(A_1 \cup A_2)$$

marks our second kind of dissonance.

The relations (6) can occur even without the occurrence of positive degrees of support for both sides of any dichotomy. In order to verify this, let us make our example numerical. Let the value of $PL(B)$ for non-empty B be given by (i) $PL(B) = 1$ if B is not contained in Honan, (ii) $PL(B) = 1/2$ if B is contained in Honan but includes points from both western and eastern Honan, and (iii) $PL(B) = 1/4$ if B is completely contained in either western or eastern Honan.

These values assure that (6) does occur, and they result in the following quantities $S(B)$ for proper subsets B of Θ: (i) $S(B) = 0$ if B does not

include the complement of Honan, (ii) $S(B)=1/2$ if B contains the complement of Honan but excludes points from both western and eastern Honan, and (iii) $S(B)=3/4$ if B contains both the complement of Honan, and all of either western Honan or eastern Honan. It is evident that S does not obey both $S(B)>0$ and $S(\bar{B})>0$ for any $B\in 2^\Theta$, for it is impossible for both B and $\bar{B}$ to contain the complement of Honan.

We should pause to remark that (6) does not require A_1 and A_2 to be disjoint. This is appropriate, for (6) is a symptom of dissonance even when A_1 and A_2 overlap. Suppose, for example, that A_1 is the western two-thirds of Honan, while A_2 is the eastern two-thirds. Then (6) would still betray the presence of some evidence pointing in both directions. In fact, it would imply the same relation for the case where A_1 and A_2 are the western and eastern halves of Honan.

We have isolated (5) and (6) as different symptoms of dissonance. In fact, however, (5) is merely a special case of (6). For if we set A_1 in (6) equal to A and A_2 equal to $\bar{A}$, we obtain

$$PL(A) < PL(A \cup \bar{A}) = 1 \quad \text{and} \quad PL(\bar{A}) < PL(A \cup \bar{A}) = 1;$$

and when this is translated by (3), it becomes (5). So (6) is the most general symptom of dissonance we have discerned so far.

1.3.2. *The Definition of Consonance*

Let us consider now a plausibility function that is completely non-dissonant. Such a function will fail to satisfy (6) and hence will satisfy

$$(7) \qquad PL(A_1 \cup A_2) = \max_{i=1,2} PL(A_i)$$

for all pairs $A_1, A_2 \in 2^\Theta$. In fact, it will satisfy

$$PL(A_1 \cup \cdots \cup A_n) = \max_{i=1,\ldots,n} PL(A_i)$$

for all finite collections $A_1, \ldots, A_n$ of elements of 2^Θ, for (7) implies (8).

Now a relation like (8) naturally causes one to ask whether the analogous relation for infinite collections should hold. Does (8) imply that

$$(9) \qquad PL\left(\bigcup_\gamma A_\gamma\right) = \sup_\gamma PL(A_\gamma)$$

for all non-empty collections $\{A_\gamma\}$ of elements of 2^Θ? Unfortunately, it does not. If Θ is infinite, then (9) is a stronger condition on PL than (8).

Nonetheless, if PL were based on strictly non-dissonant evidence, then it would be reasonable to expect it to obey (9) as well as (7) and (8). Hence I will take (9) to be the criterion for completely non-dissonant evidence. I will say that a plausibility function PL is *consonant* if it obeys (9).

One virtue of this definition is that it reduces to a much simpler form. It is easily verified that a plausibility function $PL: 2^\Theta \to [0, 1]$ is consonant if and only if

$$(10) \qquad PL(A) = \sup_{\theta \in A} PL(\{\theta\})$$

for all non-empty $A\in 2^\Theta$. This implies that a consonant plausibility function is completely determined by its values on singletons.

Conflicting evidence is the rule rather than the exception in this life, and we can expect most of the plausibility functions we meet to be dissonant. But the definition of consonance is interesting because it shows what this dissonance costs in terms of complexity. Consonant plausibility functions have a very simple structure, but dissonant ones do not, and the more dissonant they are the more complex their structure is.

1.4. The Limits of Dissonance

Though most plausibility functions exhibit some dissonance, the nature of empirical evidence seems to set limits on the degree of possible dissonance. Let us return to the example involving the map of China to see how such limits arise.

In that example, I raised the possibility that both western Honan and eastern Honan might be less plausible as places of origin of the vase than the province as a whole. If we partition the province more finely, we can adduce examples of more severe dissonance. Consider, for example, a partition into three regions as shown in Figure 6. It is possible for Honan as a whole to be more plausible than any of the regions $A_1 \cup A_2$,

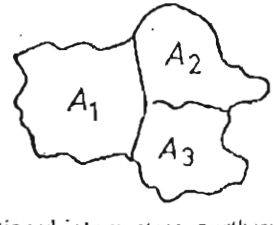

Fig. 6. Honan partitioned into western, northern and southern regions.

$A_1 \cup A_3$ or $A_2 \cup A_3$ taken singly. This would happen if for each of the three regions $A_1 \cup A_2$, $A_1 \cup A_3$ and $A_2 \cup A_3$ there were some evidence casting doubt on it but not on the remaining third of Honan.

We need not stop with thirds: given any integer n, we can partition Honan into n regions and raise the possibility that the union of any $(n-1)$ regions is subject to doubt not applying to the remainder of the providence and hence that all such unions are less plausible than the province as a whole. Following this line of thought to its furthest extreme, we might postulate that every proper subset of Honan is subject to some doubt that does not apply to all of the remainder, and hence is less plausible than the province as a whole.

Clearly, though, such extreme dissonance could never be attained on the basis of empirical evidence. Consider, for example, a subset B of Honan that falls short of including the whole province only by the exclusion of a single point. Are we to suppose that empirical evidence could cast doubt on every point of B without casting doubt on the single remaining point? One could hardly expect such precision. And it is even more farfetched that the evidence could be sufficiently voluminous and discordant for this to happen for every possible choice of the excluded point.

Surely any subset of Honan that excludes only a single point has to be accounted fully as plausible as Honan as a whole. And the same should hold true for many much smaller subsets of Honan. Consider, for example, a proper subset B that is so dense that it includes some point within an inch of every point of Honan. Such a subset B might contain only a finite number of points, but it seems hard to imagine any evidence casting doubt on it without casting doubt on all of Honan.

This example reveals two limitations of empirical evidence and two corresponding restrictions on the dissonance of plausibility functions based on empirical evidence. First, our ability to distinguish points on the map is limited, so much of the doubt cast on any point must also apply to many neighboring points. And secondly, the total amount of evidence we can acquire is limited, so our evidence cannot point in too many directions at once. Let us consider each of these limitations in turn.

1.4.1. *Topological Rules*

The mathematician will recognize the first restriction as essentially topological, and an attempt to express it precisely would lead him to adduce several rules framed in terms of the topology of Θ. Several of these rules would correspond to the regularity conditions used by Gustave Choquet to define his 'capacities'. Another rule, more easily stated but less familiar to mathematicians, would require the plausibility of any subset of Θ to equal the plausibility of its 'closure'. These topological rules are interesting, but I will not investigate them in this essay.

1.4.2. *Condensability*

The second restriction on dissonance derives not from the topology of Θ but rather from the fact that the evidence itself can have only limited complexity. There can only be so many distinct aspects of the evidence, and hence only so many distinct subsets of Θ that are touched by distinct bases of doubt. In fact, one might expect only a finite number of such fundamentally distinct aspects, or at most a countably infinite number with a finite number predominating in importance.

Fixing our attention on a given subset A of Θ, we can adduce similar considerations relative to the bases of doubt that apply to proper subsets of A but not to all of A. If there were only a finite number of these, then by choosing a point of A immune to each we could form a finite subset B of A which, as a whole, was immune to all the bases of doubt. This finite subset $B \subset C$ would be fully as plausible as A itself. If, on the other hand, there were an infinite number of bases of doubt, a finite number of which carried most of the weight, then we could expect to find a finite subset $B \subset A$ which was nearly as plausible as A.

This second restriction leads us, then, to the rule that

$$(11) \qquad PL(A) = \sup \{PL(B) \mid B \subset A\colon B \text{ is finite}\}$$

for all $A \in 2^{\Theta}$. A plausibility function that obeys this rule is called *condensable* – a name based on the intuitive idea that at least most of the plausibility of A can be 'condensed' onto a finite subset.

The considerations just adduced are meant to explain the intuitive significance of condensability; they hardly constitute a demonstration that plausibility functions ought always to be condensable. I have found, however, that condensability is characteristic of empirical evidence, and it plays an important role in the theory developed in this essay. As will

see in §3, condensability is one of the conditions that must be met if Dempster's rule of combination is to be fully applicable.

It is also interesting to note that the criterion for consonance, (9), is equivalent to (7) once condensability is assumed.

2. Statistical evidence

I have explained at length what degrees of support would look like if we could calculate them, but it remains painfully obvious that we usually cannot. In most cases where we want to assess the evidence for a proposition there is simply no quantitative structure that can be exploited to produce numbers. But we find an exception in *statistical evidence* – a type of evidence that has so rich a quantitative structure that the calculation of numerical degrees of support is quite conceivable.

2.1. The problem of statistical support

The notion of statistical evidence depends on the notion of an aleatory law, or an objective probability law. An aleatory law is a law that tells an experiment's propensity for producing each of its various possible outcomes. The first step in specifying an aleatory law is to specify the set X of all possible outcomes of the experiment. If X is conceived of as a topological space, then the further description of the aleatory law is somewhat complicated. But in the case where X is 'discrete' it is quite simple. One simply specifies, for each $x \in X$, a quantity $P(x)$ which is the experiment's propensity for producing the outcome x. The quantity $P(x)$ is also called x's objective probability, and it is the frequency with which x will occur in a long sequence of physically independent trials of the experiment. Each of the quantities $P(x)$, $x \in X$, is non-negative, and they add to one. In the following discussion I will concern myself mainly with the case where X is discrete.

Let us return to the problem of support for a parameter θ whose possible values constitute a set Θ. Suppose that θ is related to a given experiment by a *statistical specification* $\{P_\theta\}_{\theta \in \Theta}$. This means that to each element $\theta \in \Theta$ there corresponds an aleatory law P_θ on X, and that the experiment is actually governed by the aleatory law corresponding to the true value of θ. Since the different aleatory laws may attribute different propensities or objective probabilities to the various possible outcomes

in X, the outcomes that are observed in a sequence of physically independent trials will constitute evidence as to which aleatory law actually governs the process and hence as to which element of Θ is the true value of θ. It is reasonable to call this type of evidence *statistical evidence*. If it is the only kind of evidence we have about θ, then the problem of measuring degrees of support for θ is a *problem of statistical support*.

It is useful to distinguish between statistical specifications that are complete and those that are restricted. A *complete* specification $\{P_\theta\}_{\theta \in \Theta}$ on X is one that includes every possible aleatory law on X; a *restricted* one is one that is not complete. (This terminology is not standard.) Obviously, every restricted specification on X can be thought of as a subset of the essentially unique complete specification on X.

The idea of using the outcomes of an experiment as evidence about which of a class $\{P_\theta\}_{\theta \in \Theta}$ of aleatory laws governs it is a familiar one. I should enlarge, though, on what is meant when an aleatory law P is said to govern an experiment. This is taken to mean not only that the experiment's propensity to produce the outcome x in a single trial is $P(x)$, but also that its propensity to produce the sequence $(x_1, \ldots, x_n)$ of outcomes in a sequence of physically independent trials is $P(x_1) \ldots P(x_n)$.

2.1.1. *Dissonance*

I just asserted that the outcomes observed in a sequence of trials of an experiment constitute a body of evidence about which aleatory law governs the experiment. The thought behind this assertion is that the observation of an outcome $x \in X$ is evidence in favor of those laws that attribute to the experiment the greatest propensity for producing x. More generally, it tends to favor any law that attributes a given probability to x over a law that attributes a smaller probability to x.

The straightforward appearance of the evidence provided by a single observation x might lead us to think of it as highly consonant evidence. After all, it points in just one direction – towards those aleatory laws that attribute a high probability to x. In some cases, however, there will be several quite different aleatory laws that attribute a high probability to x, and since the evidence points towards each of these, it might be said to point in many directions at once. Hence it is not clear that the evidence provided by a single observation will always be consonant; in some cases it might be better to think of it as dissonant.

While there may often be some question about the consonance or dissonance of the evidence in the case of a single observation, there will usually be little question in the case of several observations. In that case, the evidence will almost always be highly dissonant. For even if each single observation points in a single direction, the different observations will most likely point in different directions.

So we must consider two possible types of dissonance: dissonance arising from the combination of observations and dissonance arising from a single observation.

2.1.2. *Dissonance from the Combination of Observations*

Let us consider the simplest of statistical specifications: the *binomial specification*. Suppose $X=\{\text{Heads, Tails}\}$ and $\Theta=[0, 1]$, with $P_\theta(\text{Heads})=\theta$ $P_\theta(\text{Tails})=1-\theta$. In other words, θ is a coin-tossing experiment's propensity or probability for producing heads, and θ might have any value between zero and one. Obviously, a flip resulting in tails will be evidence for a low value. Any single flip by itself will be quite straightforward evidence, but a flip resulting in heads will point in the direction opposite to a flip resulting in tails. So while we may expect a single observation x to produce a consonant support function, we must expect a sequence $\mathbf{x}=(x_1, \ldots, x_n)$ of observations to produce a dissonant support function, at least if $\mathbf{x}$ includes both some heads and some tails. In other words, we might expect the support function S_{x_i} arising from the ith trial to be consonant, but we would expect the support function $S_{\mathbf{x}}$ resulting from the combination of the support functions $S_{x_1}, \ldots, S_{x_n}$ to be dissonant.

2.1.3. *Dissonance from a Single Observation*

Suppose we are confronted with a closed box of recently hatched chicks. We know that the chicks include both Rhode Island Reds and White Leghorns, and both males and females, but we are uncertain of the proportions in the different categories. There may be equal numbers of each breed, but either more males or fewer males than females. Or there may be equal numbers of each sex, but either more Rhode Island Reds or fewer Rhode Island Reds than White Leghorns. For the sake of concreteness, let us suppose that there are only four possibilities: (i) equal numbers of each breed but four times as many males as females, (ii) equal numbers of each breed but four times as many females as males, (iii) equal numbers of each sex but four times as many Reds as Whites, or (iv) equal numbers of each sex but four times as many Whites as Reds. Denote these four possibilities by θ_M, θ_F, θ_R and θ_W, respectively, and set $\Theta=\{\theta_M, \theta_F, \theta_R, \theta_W\}$.

Now suppose we draw a chick from the box 'at random'. Then the element of Θ that correctly represents the contents of the box will determine an aleatory law that will govern the result of the draw. That aleatory law will be defined, of course, on the set $X=\{\text{Red Female, Red Male, White Female, White Male}\}$. The four possible aleatory laws are shown in Table VII.

TABLE VII

The four aleatory laws

	Red Female	Red Male	White Female	White Male
θ_F	0.4	0.1	0.4	0.1
θ_M	0.1	0.4	0.1	0.4
θ_R	0.4	0.4	0.1	0.1
θ_W	0.1	0.1	0.4	0.4

Finally, suppose our single draw results in a Red female chick. What sort of evidence does this provide as to the correct aleatory law?

Clearly, the evidence points both towards θ_F and towards θ_R. The fact that a female chick was drawn points towards θ_F; and the fact that a Red check was drawn points towards θ_R. So the result of the single draw points in two different directions at once; it provides dissonant evidence.

2.2. THE FIRST POSTULATE OF PLAUSIBILITY

In a problem of statistical support, the propositions about θ are sometimes called hypotheses, and one distinguishes between simple hypotheses and composite hypotheses. A subset $\{\theta\}\subset\Theta$ that contains a single element θ is a simple hypothesis; it asserts that the experiment is governed by the aleatory law P_θ. A subset of Θ that has more than one element, on the other hand, is a composite hypothesis; it asserts only that the experiment is governed by one of several aleatory laws. As it turns out, it is easier to investigate the plausibilities of simple hypotheses than the plausibilities of composite hypotheses.

Suppose, indeed, that one has observed the outcome x in a trial of the experiment and wants to compare the plausibilities of the two simple hypotheses $\{\theta_1\}$ and $\{\theta_2\}$. Is it more plausible that the experiment is governed by P_{θ_1} or that it is governed by P_{θ_2}? We answer this question, of course, by comparing $P_{\theta_1}(x)$ and $P_{\theta_2}(x)$: the simple hypothesis that attributes the greater objective probability to the actual observation x will be the more plausible one.

It seems reasonable to go even farther and postulate that the degree of plausibility of a simple hypothesis $\{\theta\}$ should be proportional to the quantity $P_\theta(x)$. Denoting by PL_x the plausibility function on 2^Θ resulting from the observation x, this postulate can be written in symbols as

$$PL_x(\{\theta\}) = c(x)\, P_\theta(x),$$

where $c(x)$ depends on x but not on θ.

For the sake of economy, I will state this postulate formally for the case where $\{P_\theta\}_{\theta\in\Theta}$ is complete:

(I) Suppose $\{P_\theta\}_{\theta\in\Theta}$ is a complete statistical specification on the discrete space X, and suppose $PL_x\colon 2^\Theta \to [0, 1]$ is the plausibility function based on the single observation $x\in X$. Then

$$PL_x(\{\theta\}) = c(x)\, P_\theta(x),$$

where $c(x)$ depends on x but not on θ.

This is the *first postulate of plausibility.*

In the presence of other postulates that I will adopt in §4, this postulate implies the following more general statement, which applies to both complete and restricted specifications and to any number of observations:

(I′) Suppose $\{P_\theta\}_{\theta\in\Theta}$ is a statistical specification on the discrete space X, and suppose $PL_x\colon 2^\Theta \to [0, 1]$ is the plausibility function based on the observations $\mathbf{x}=(x_1, \ldots, x_n)$. Then

$$PL_{\mathbf{x}}(\{\theta\}) = c(\mathbf{x})\, P_\theta(x_1)\ldots P_\theta(x_n),$$

where $c(\mathbf{x})$ depends on $\mathbf{x}$ but not on θ.

The first postulate of plausibility is hardly a novel idea. Every statistician will agree that the objective probability that a simple hypothesis atrributes to the actual observations is a measure of the simple hypothesis' 'relative plausibility', 'likelihood', 'probability', or some such thing. The idea can be traced back at least to Johann Heinrich Lambert's *Photometria*, published in 1760. Daniel Bernoulli toyed with the idea about the same time, and shortly later it was incorporated into the 'Bayesian' framework that was firmly imposed on statistics by the consummate politician named Pierre Simon Laplace. In this century it was forcefully re-extracted from that framework by R. A. Fisher, who called the quantity $P_\theta(x)$, considered as a function of θ, the 'likelihood' of θ.

The aspect of the present formulation that is novel is the explicit recognition that it is only the plausibilities of simple hypotheses that are proportional to the objective probabilities. The first postulate tells us only about the quantities $PL_x(\{\theta\})$ for $\theta\in\Theta$, and when the evidence is dissonant these will not determine the quantities $PL_x(A)$ for composite hypotheses $A\in 2^\Theta$.

I will now consider the first postulate of plausibility in the light of two examples, one involving a complete specification, and the other involving a restricted specification.

2.2.1. *The Binomial Specification*

Consider first the example mentioned earlier, where $X=\{\text{Head, Tails}\}$, $\Theta=[0, 1]$, and P_θ (Heads)$=\theta$, P_θ (Tails)$=1-\theta$. This specification is complete, and since X has two elements it is called a *binomial* specification.

Suppose we have six observations, $\mathbf{x}=(x_1, \ldots, x_6)$ from this specification, and three of them are heads while the other three are tails. Then by the first postulate of plausibility,

$$\begin{aligned} PL_{\mathbf{x}}(\{\theta\}) &= c(\mathbf{x})\, P_\theta(x_1)\ldots P_\theta(x_6) \\ &= c(\mathbf{x})\, [P_\theta(\text{Heads})]^3\, [P_\theta(\text{Tails})]^3 \\ &= c(\mathbf{x})\, \theta^3(1-\theta)^3. \end{aligned}$$

Since $\theta^3(1-\theta)^3$ takes its maximum value when $\theta=1/2$, $1/2$ will be the most plausible single value for θ. Any other value θ will have a degree of plausibility that is only $64\theta^3(1-\theta)^3$ as great as the degree of plausibility for $1/2$.

But this is all the first postulate of plausibility tells us. It does not tell us the absolute value of $PL_x\{1/2\}$ or of $PL_x(\{\theta\})$ for any other $\theta\in\Theta$. And it tells us nothing about the degrees of plausibility for composite hypotheses.

2.2.2. *A Restricted Binomial Specification*

Now let us consider the restricted statistical specification that is obtained from the preceding example by taking Θ to be the pair $\{1/3, 2/3\}$ instead of the whole interval $[0, 1]$. In other words, let us suppose that θ, the coin's propensity for coming up heads, is known to have either the value 1/3 or the value 2/3. Still assuming that we have obtained three heads and three tails, the plausibilities for the simple hypotheses $\{1/3\}$ and $\{2/3\}$ will be

$$PL_x(1/3) = c(x)\,(1/3)^3\,(2/3)^3$$

and

$$PL_x(2/3) = c(x)\,(2/3)^3\,(1/3)^3\,.$$

In other words, they will be the same. But how great will their common value be?

Both alternatives are equally plausible, and they are the only alternatives, so one might be inclined to award them both plausibility one. But this would correspond to saying that the six tosses have really produced no evidence at all. In fact, they have produced conflicting evidence; each of the two values has had some doubt cast on it as a result of the six tosses. Doubt was cast on the value 1/3 every time heads came up, and doubt was cast on the value 2/3 every time tails came up.

Another way to see that our evidence is internally conflicting rather than null is to observe that it has an effect when it is combined with further evidence. Suppose, for example, that another six tosses of the coin result in four heads and two tails. Now this new evidence, considered on its own, should produce a mild degree of support and a higher degree of plausibility for $\{2/3\}$. But when we combine it with the old evidence, we end up with seven heads out of twelve tosses. This overall result still lends more support to $\{2/3\}$ than to $\{1/3\}$, but it surely casts more doubt on $\{2/3\}$ than did the observation of four heads out of six tosses. So the three heads and three tails do differ from no evidence when combined with more tosses.

We can describe this phenomenon more generally. The effect of the three heads and three tails will be to counter any more one-sided evidence arising from further tosses. And the same countering effect will result from any equal number of heads and tails – the greater the number, the stronger the effect. If, for example, our initial evidence consists of fifty heads out of a hundred tosses, then it will practically nullify the more one-sided evidence provided by four heads and six tosses.

So the common degree of plausibility resulting from three heads and three tails must be less than one. On the other hand, it must be greater than one-half, for the two plausibilities must obey

$$PL_x(\{1/3\}) + PL_x(\{2/3\}) = PL_x(\{1/3\}) + PL_x(\overline{\{1/3\}}) \geq 1\,.$$

The theory of §4 below gives the value 0.6, resulting in the dissonant plausibility function in Table VIII.

TABLE VIII

S_x and PL_x when x consists of three heads and three tails.

A	$S_x(A)$	$PL_x(A)$
0	0	0
{1/3}	0.4	0.6
{2/3}	0.4	0.6
Θ	1	1

2.3. THE SECOND POSTULATE OF PLAUSIBILITY

The first postulate of plausibility concerns the plausibilities of subsets of Θ that consist of single elements. As a first step beyond the first postulate, it is natural to consider subsets of Θ that consist of two elements.

Consider a doubleton $\{\theta_1, \theta_2\} \in 2^\Theta$, and suppose we have a single observation $x \in X$. If PL_x were consonant, we would have

$$PL_x(\{\theta_1, \theta_2\}) = \max_{i=1,2} PL_x(\{\theta_i\})\,.$$

But if PL_x were dissonant with respect to the pair θ_1, θ_2, we would have

$$(12) \qquad PL_x(\{\theta_1, \theta_2\}) > \max_{i=1,2} PL_x(\{\theta_i\})\,.$$

When ought (12) to occur? In other words, when is the evidence x dissonant with respect to θ_1, θ_2?

We may assume that $P_{\theta_1}(x) \geq P_{\theta_2}(x)$, so that $PL_x(\{\theta_1\}) \geq PL_x(\{\theta_2\})$ by

the first postulate. Our question then becomes whether we ought to have

$$PL_x(\{\theta_1, \theta_2\}) > PL_x(\{\theta_1\}) \tag{13}$$

even though $PL_x(\{\theta_1\}) \geq PL_x(\{\theta_2\})$. In other words, when ought the addition of θ_2 to the hypothesis $\{\theta_1\}$ cause an increase in plausibility even though P_{θ_2} attributes no greater objective probability to the actual observation than P_{θ_1}?

This is a difficult question. On the whole, of course, the evidence x points towards θ_1 more strongly than towards θ_2. (Or equally strongly if $P_{\theta_1}(x) = P_{\theta_2}(x)$.) But if some aspect of the evidence x points more towards θ_2 than towards θ_1, we will have an example of dissonance, and (13) should hold.

Without answering definitely the question of when (13) should hold, we can specify one situation where nearly everyone would agree that it should not hold. Suppose that for every $y \in X$,

$$\frac{P_{\theta_1}(x)}{P_{\theta_1}(y)} \geq \frac{P_{\theta_2}(x)}{P_{\theta_2}(y)}. \tag{14}$$

This is stronger than saying that P_{θ_1} attributes a greater probability to x than P_{θ_2} does. For it says that for every other possible outcome y, P_{θ_1} attributes a greater probability to x relative to y than P_{θ_2} does. If (14) holds, then it would seem that every aspect of the evidence points towards θ_1 more strongly than towards θ_2. Certainly, the comparison of the actual outcome with any other possible outcome favors θ_1 over θ_2. Hence we may conclude that (13) should not hold if (14) holds.

We have arrived at the second postulate. As in the case of the first postulate, I will state it formally for the case where the specification is complete:

(II) Suppose $\{P_\theta\}_{\theta \in \Theta}$ is a complete statistical specification on the discrete space X, and suppose $PL_x: 2^\Theta \to [0, 1]$ is the plausibility function based on the single observation $x \in X$. Then

$$PL_x(\{\theta_1, \theta_2\}) = PL_x(\{\theta_1\})$$

whenever θ_1 and θ_2 are elements of Θ and

$$\frac{P_{\theta_1}(x)}{P_{\theta_1}(y)} \geq \frac{P_{\theta_2}(x)}{P_{\theta_2}(y)}$$

for all $y \in X$.

This is the *second postulate of plausibility.*

In the presence of the further postulates adopted in §4, this postulate implies the following more general statement:

(II′) Suppose $\{P_\theta\}_{\theta \in \Theta}$ is a statistical specification on a discrete space X, and suppose $PL_x: 2^\Theta \to [0, 1]$ is the plausibility function based on the observations $\mathbf{x} = (x_1, \ldots, x_n)$. Then

$$PL_x(\{\theta_1, \theta_2\}) = PL_x(\{\theta_1\})$$

whenever θ_1 and θ_2 are elements of Θ and

$$\frac{P_{\theta_1}(x_1) \ldots P_{\theta_1}(x_n)}{P_{\theta_1}(y_1) \ldots P_{\theta_1}(y_n)} \geq \frac{P_{\theta_2}(x_1) \ldots P_{\theta_2}(x_n)}{P_{\theta_2}(y_1) \ldots P_{\theta_2}(y_n)}$$

for all sequences $\mathbf{y} = (y_1, \ldots, y_n)$ of elements of X.

2.3.1. *The Binomial: One Observation*

Consider the complete binomial specification: $X = \{\text{Heads, Tails}\}$, $\Theta = [0, 1]$, and $P_\theta(\text{Heads}) = \theta$, $P_\theta(\text{Tails}) = 1 - \theta$. And suppose we have a single observation $x = \text{Heads}$. Then (14) becomes

$$\frac{P_{\theta_1}(\text{Heads})}{P_{\theta_1}(y)} \geq \frac{P_{\theta_2}(\text{Heads})}{P_{\theta_2}(y)}$$

for all $y \in X$. But this means that

$$\frac{\theta_1}{\theta_1} \geq \frac{\theta_2}{\theta_2} \quad \text{and} \quad \frac{\theta_1}{1 - \theta_1} \geq \frac{\theta_2}{1 - \theta_2},$$

and this is equivalent to $\theta_1 \geq \theta_2$. Hence the second postulate says in this case that

$$PL_{\text{Heads}}(\{\theta_1, \theta_2\}) = PL_{\text{Heads}}(\{\theta_1\})$$

whenever $\theta_1 \geq \theta_2$.

Hence

$$PL_{\text{Heads}}(\{\theta_1, \theta_2\}) = \max_{i = 1, 2} PL_{\text{Heads}}(\{\theta_i\})$$

for all pairs θ_1, θ_2; there is no pair θ_1, θ_2 for which dissonance is exhibited. This reflects the consonant nature of the evidence consisting of a single outcome of heads. Such evidence points unambiguously towards the aleatory laws that attribute greater probability to heads.

Notice in particular that

$$PL_{\text{Heads}}(\{1, 0\}) = PL_{\text{Heads}}(\{1\})$$

for all $0 \in [0, 1]$; there is no aleatory law whose addition will increase the plausibility of the simple hypothesis that attributes probability one to heads.

2.3.2. *The Binomial: Many Observations*

While discussing the first postulate, I considered an example of six observations $\mathbf{x} = (x_1, \ldots, x_6)$ from the restricted binomial specification $\Theta = \{1/3, 2/3\}$. Assuming that $\mathbf{x}$ consisted of three heads and three tails, I obtained the dissonant plausibility function in Table VIII. Is the dissonance in that plausibility function permitted by (II')?

The plausibility function $PL_{\mathbf{x}}$ in Table VIII is dissonant because

$$PL_{\mathbf{x}}(\{1/3, 2/3\}) > PL_{\mathbf{x}}(\{1/3\}) = PL_{\mathbf{x}}(\{2/3\}).$$

This is permitted by our second postulate only if

$$(15) \qquad \frac{P_{1/3}(x_1) \ldots P_{1/3}(x_6)}{P_{1/3}(y_1) \ldots P_{1/3}(y_6)} < \frac{P_{2/3}(x_1) \ldots P_{2/3}(x_6)}{P_{2/3}(y_1) \ldots P_{2/3}(y_6)}$$

holds for some $\mathbf{y}$. But (15) will indeed hold if we choose a $\mathbf{y}$ consisting, say, of six tails. For then (15) will become

$$\frac{(1/3)^3 (2/3)^3}{(2/3)^6} < \frac{(2/3)^3 (1/3)^3}{(1/3)^6},$$

or $(1/2)^3 < 2^3$. Hence the second postulate does allow the dissonance in this example. This can be explained by pointing out that while $\{2/3\}$ attributes no greater likelihood to the overall observations $\mathbf{x}$ than $\{1/3\}$ does, it does attribute a greater likelihood to $\mathbf{x}$ relative to a sequence $\mathbf{y}$ containing even more tails.

2.3.3. *A Trinomial Example*

Set $X = \{\text{Azure, Brown, Crimson}\}$, and let $\{P_\theta\}_{\theta \in \Theta}$ be the complete statistical specification on X. This specification is most easily described by setting.

$$\Theta = \{(a, b, c) \mid a \geq 0, b \geq 0, c \geq 0; a + b + c = 1\}$$

and setting $P_\theta(\text{Azure}) = a$, $P_\theta(\text{Brown}) = b$, and $P_\theta(\text{Crimson}) = c$ when $\theta = (a, b, c)$.

In order to think about the second postulate, consider the following elements of Θ:

$$\theta_1 = (3/4, 1/3, 1/8) \qquad \theta_3 = (1/2, 1/6, 1/3)$$
$$\theta_2 = (1/2, 1/4, 1/4) \qquad \theta_4 = (1/4, 0, 3/4)$$
$$\theta_5 = (1, 0, 0)$$

Now suppose we have a single observation $x = \text{Azure}$. Then what does the second postulate tell us about the plausibilities of the various doubletons that can be formed from these five elements of Θ?

Well, the second postulate will require that

$$PL_{\text{Azure}}(\{\theta_i, \theta_j\}) = PL_{\text{Azure}}(\{\theta_i\})$$

whenever

$$\frac{P_{\theta_i}(\text{Azure})}{P_{\theta_i}(y)} \geq \frac{P_{\theta_j}(\text{Azure})}{P_{\theta_j}(y)}$$

for all $y \in X$: i.e.; whenever

$$\frac{P_{\theta_i}(\text{Azure})}{P_{\theta_i}(\text{Brown})} \geq \frac{P_{\theta_j}(\text{Azure})}{P_{\theta_j}(\text{Brown})} \qquad \text{and} \qquad \frac{P_{\theta_i}(\text{Azure})}{P_{\theta_i}(\text{Crimson})} \geq \frac{P_{\theta_j}(\text{Azure})}{P_{\theta_j}(\text{Crimson})}.$$

This means that the second postulate will require

$$PL_{\text{Azure}}(\{\theta_1, \theta_2\}) = PL_{\text{Azure}}(\{\theta_1\}),$$
$$PL_{\text{Azure}}(\{\theta_1, \theta_3\}) = PL_{\text{Azure}}(\{\theta_1\}),$$

and

$$PL_{\text{Azure}}(\{\theta_5, \theta_i\}) = PL_{\text{Azure}}(\{\theta_5\})$$

for $i = 1, \ldots, 4$. But it will *not* require

$$PL_{\text{Azure}}(\{\theta_2, \theta_3\}) = PL_{\text{Azure}}(\{\theta_2\}),$$

nor

$$PL_{\text{Azure}}(\{\theta_1, \theta_4\}) = PL_{\text{Azure}}(\{\theta_1\}).$$

In other words, it will allow

$$(16) \qquad PL_{\text{Azure}}(\{\theta_2, \theta_3\}) > PL_{\text{Azure}}(\{\theta_2\})$$

and

$$(17) \qquad PL_{\text{Azure}}(\{\theta_1, \theta_4\}) > PL_{\text{Azure}}(\{\theta_1\}),$$

even though $PL_{\text{Azure}}(\{\theta_3\}) = PL_{\text{Azure}}(\{\theta_2\})$ and $PL_{\text{Azure}}(\{\theta_4\}) < PL_{\text{Azure}}(\{\theta_1\})$.

Notice that if (16) and (17) actually hold then this will be an example in which dissonance is displayed even though there is but a single observation. This may seem strange, for the observation $x = \text{Azure}$ might be thought to point unambiguously towards those aleatory laws which attribute the greater probability to Azure. Actually, it does point towards the aleatory laws attributing a greater probability to Azure, but this does not define an unambiguous 'direction'. In comparing θ_2 and θ_3, for example, we see that θ_2 attributes the greater likelihood to Azure relative to Brown, while θ_3 attributes the greater likelihood to Azure relative to Crimson.

2.3.4. *The Geometry of the Trinomial Specification*

The application of the second postulate to the trinomial specification can be understood more easily if the set Θ is represented by an equilateral triangle.

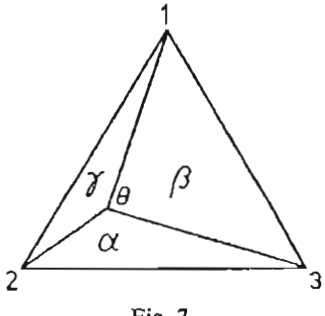

Fig. 7.

As we see in Figure 7, any point θ of an equilateral triangle divides the triangle into three smaller triangles α, β, and γ, each based on one of the three sides of the equilateral triangle. Let us suppose that the total area of the triangle is equal to one, and denote the areas of the triangles α, β and γ by a, b and c, respectively. Then $a+b+c=1$. The triplet (a, b, c) is called the *barycentric coordinates* of the point θ. As the position of

θ varies, the areas a, b and c will vary; and it is evident that by placing θ in the right place they can be made to assume any triplet of non-negative values adding to one. Hence the points θ of the triangle are in a one-to-one correspondence with the elements θ of Θ.

Now consider a ray emanating from vertex 3, as in Figure 8, and consider any two points $\theta_1 = (a_1, b_1, c_1)$ and $\theta_2 = (a_2, b_2, c_2)$ on that ray. It is easily seen that $a_1/b_1 = a_2/b_2$. In other words, the ratio a/b is constant for points on a given ray from vertex 3. Furthermore, that ratio increases as the ray is raised; in Figure 9, for example, the ratio is higher for θ_1 than for θ_2.

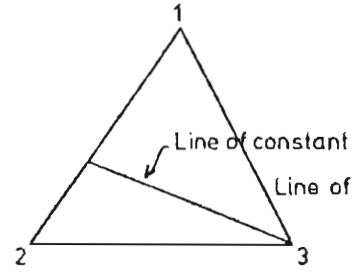

Fig. 8.

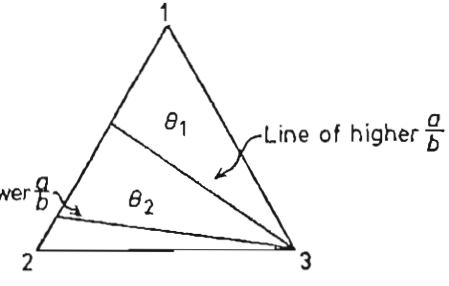

Fig. 9.

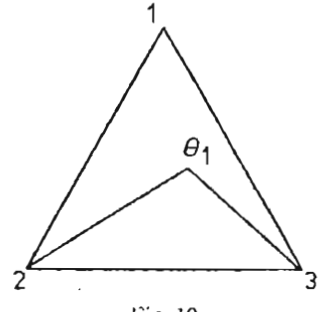

Fig. 10.

Similarly, rays from vertex 2 are lines of constant a/c; the higher the ray the higher that constant.

Now as we saw above, the plausibility function PL_{Azure} is required by the second postulate to obey $PL_{\text{Azure}}(\{\theta_1, \theta_2\}) = PL_{\text{Azure}}(\{\theta_1\})$ whenever both

$$(18) \qquad \frac{a_1}{b_1} \geqslant \frac{a_2}{b_2} \quad \text{and} \quad \frac{a_1}{c_1} \geqslant \frac{a_2}{c_2},$$

where $\theta_1=(a_1, b_1, c_1)$ and $\theta_2=(a_2, b_2, c_2)$. And it is easy to see that (18) holds only when θ_2 is inside the triangle $\theta_1 23$. If θ_2 is outside that triangle, then the second postulate will permit $PL_{\text{Azure}}(\{\theta_1, \theta_2\}) > > PL_{\text{Azure}}(\{\theta_1\})$.

3. Belief

Our theory of statistical evidence seems promising, but it has reached an impasse. I have adduced several general rules for support functions, and I have added two postulates that are specific to statistical evidence. Yet all of these conditions together fail to determine a unique real function $S_x: 2^\Theta \to [0, 1]$ that can be regarded as measuring the degrees to which the statistical evidence x supports the different statistical hypotheses in 2^Θ. We evidently need yet further conditions on the function S_x.

I propose to search for some of these conditions in a theory of partial belief. I am going to adduce general rules governing a function Bel: $2^\Theta \to [0, 1]$ that purports to measure a person's degrees of belief in the various propositions in 2^Θ, and then I am going to require that the support functions S_x also obey these rules. As we will see in §4, this requirement will still fail to determine the support functions S_x uniquely, but it will considerably narrow the range of possibilities.

Why do I want to cross our theory of evidence with a theory of partial belief? Since the degree of support for a proposition ought to determine one's degree of belief in it, I might excuse myself by claiming that degrees of support should obey rules applicable to degrees of belief. But this would be at most a partial justification; for the fact that the degree of support ought to determine one's degree of belief hardly implies that the first should obey any rule applicable to the second. The real reason for turning to a theory of partial belief lies in the differences rather than in the similarities between the notions of support and belief. The fact is that the subjective notion of partial belief has a stronger structure than the logical notion of partial support, so that it is possible to adduce rules for degrees of belief that cannot be adduced so naturally for degrees of support. Indeed, it is natural to think not only that partial belief is related to complete belief as a part is to a whole, but also that the partial beliefs accorded to different propositions correspond to parts of the same whole. This amounts to reifying 'our total belief' – to thinking of it as a fixed substance that can be divided up in various ways. It is not nearly so natural to reify 'the total support' in this way.

This section outlines a theory based on the notion that we have a certain degree of belief in a proposition when we have committed to it that proportion of our total belief – or if you prefer, that proportion of our total 'probability'. A fuller account of this theory can be found in my *Allocations of Probability: A Theory of Partial Belief*.

3.1. Belief functions

A function $\text{Bel}: 2^\Theta \to [0, 1]$ purporting to give a person's degrees of belief is called a *belief function* if it obeys these three axioms:

(I) $\quad \text{Bel}(\emptyset) = 0.$

(II) $\quad \text{Bel}(\Theta) = 1.$

(III) $\quad \text{Bel}(A_1 \cup \ldots \cup A_n) \geq \sum_i \text{Bel}(A_i) - \sum_{i<j} \text{Bel}(A_i \cap A_j) + \ldots + (-1)^{n+1} \text{Bel}(A_1 \cap \ldots \cap A_n)$

whenever

$$A_1, \ldots, A_n \in 2^\Theta.$$

These axioms imply all the rules for support functions listed in Table I. Hence any belief function will formally qualify as a support function. On the other hand, these axioms are more restrictive than the rules for support functions, and hence not every function satisfying those rules necessarily qualifies as a belief function. Nevertheless, all the support functions for the case of two alternatives – i.e., all the support functions described in Table V – do qualify as belief functions. Other support functions that qualify as belief functions include the vacuous ones, all consonant ones, and many dissonant ones.

3.1.1. *Deriving the Axioms*

The axioms for belief functions follow naturally from the idea that $\text{Bel}(A)$, as our degree of belief in A, is the measure of that portion of our belief that is committed to A.

Axioms (I) and (II) are unexceptionable; we already adduced them for support functions. But here they acquire a stronger intuitive meaning. The statement that $\text{Bel}(\emptyset)=0$ reflects the fact that none of our belief

ought to be committed to the impossible proposition $\emptyset$. And the statement that $\mathrm{Bel}(\Theta)=1$ reflects the fact that all of our belief ought to be committed to the sure proposition Θ and the convention that the measure of our total belief is equal to one.

Axiom (III) looks a little more formidable. To begin with, it is really an infinite number of axioms, one for each natural number n. Before asking you to swallow it whole, I will ask you to consider some of its simpler consequences. For $n=2$, the axiom becomes

$$(19) \qquad \mathrm{Bel}(A_1 \cup A_2) \geq \mathrm{Bel}(A_1) + \mathrm{Bel}(A_2) - \mathrm{Bel}(A_1 \cap A_2)$$

for all pairs A_1, A_2 of subsets of Θ. Now when A_1 and A_2 are disjoint, or $A_1 \cap A_2 = \emptyset$, $\mathrm{Bel}(A_1 \cap A_2)=0$. So one consequence of (19) is:

$$(20) \qquad \text{If } A_1 \cap A_2 = \emptyset, \text{ then } \mathrm{Bel}(A_1 \cup A_2) \geq \mathrm{Bel}(A_1) + \mathrm{Bel}(A_2).$$

Now suppose $A \subset B$ and set $A_2 = B - A$. Then $B = A \cup A_2$ and $A \cap A_2 = \emptyset$. Hence (20) will give $\mathrm{Bel}(B) = B(A \cup A_2) \geq \mathrm{Bel}(A) + \mathrm{Bel}(A_2)$. And hence $\mathrm{Bel}(B) \geq \mathrm{Bel}(A)$. So one consequence of (20) is our familiar rule of monotonicity:

$$(21) \qquad \text{If } A \subset B, \text{ then } \mathrm{Bel}(A) \leq \mathrm{Bel}(B).$$

Let us develop the intuitive arguments for (19), (20), and (21), beginning with (21) and working backwards.

I would argue for (21) as follows: Since $A \subset B$, the proposition A implies the proposition B. Hence any belief I commit to A I must also commit to B; and the total portion I commit to B will therefore include the total portion I commit to A. And hence $\mathrm{Bel}(B)$, the measure of the total portion of belief committed to B, will be at least as great as $\mathrm{Bel}(A)$, the measure of the total portion of belief committed to A.

The defense of (20) is similar. First we note that any belief committed to A_1 must also be committed to $A_1 \cup A_2$, since $A_1 \subset A_1 \cup A_2$. Similarly, any belief committed to A_2 must also be committed to $A_1 \cup A_2$. And there can be no overlap between the belief committed to A_1 and the belief committed to A_2; the relation $A_1 \cap A_2 = \emptyset$ means that as propositions A_1 and A_2 are incompatible, and a single portion of belief can hardly be committed to both of two incompatible propositions. Hence the total belief committed to $A_1 \cup A_2$ will include the two disjoint portions that are committed to A_1 and A_2, respectively; and its measure will be at least as great as the sum of the measures of these two disjoint portions.

When the two propositions A_1 and A_2 are compatible (i.e., $A_1 \cap A_2 \neq \emptyset$), there may be some overlap between the portion of belief committed to A_1 and the portion of belief committed to A_2. In fact, the overlap – or the belief that is committed both to A_1 and to A_2 – will consist precisely of the belief that is committed to $A_1 \cap A_2$. This fact provides the basis for (19). For it means that the quantity $\mathrm{Bel}(A_1) + \mathrm{Bel}(A_2) - \mathrm{Bel}(A_1 \cap A_2)$ measures the total belief that is committed either to A_1, to A_2 or to both; and all that belief must be included in the total belief committed to $A_1 \cup A_2$, which is measured by $\mathrm{Bel}(A_1 \cup A_2)$.

The versions to Axiom (III) for other values of n can be justified by similar, but progressively more convoluted arguments.

3.1.2. *Allocations of Probability*

The axioms for belief functions are based on an intuitive picture wherein various portions of our belief, or various of our probability masses, are committed to various propositions. This intuitive picture can be made more precise by the notion of an *allocation of probability*.

The first step in formalizing the intuitive picture is the assumption that the set M of probability masses has the mathematical structure of a *complete Boolean algebra*. Let me outline roughly what this means. First we suppose that for every pair M_1, M_2 of probability masses in M there is another probability mass $M_1 \vee M_2$ in M, which is their 'union', and yet another, $M_1 \wedge M_2$, which is their overlap or 'intersection'. Of course, M_1 and M_2 may be disjoint, in which case the probability mass $M_1 \wedge M_2$ will be null – there will be no probability in it. We can use the symbol Λ to represent the null probability mass and the symbol V to represent the probability mass consisting of all our probability. Besides unions and intersections for pairs we also require unions and intersections for larger collections of probability masses. For any collection $\{M_\gamma\}$ of elements of M we require the existence of a union $\bigvee M_\gamma$ and an intersection $\bigwedge M_\gamma$. And for each $M \in \mathsf{M}$ we require the existence of a probability mass in M that consists precisely of all the probability not in M. This probability mass is called the complement of M and is denoted $\bar{M}$; M and $\bar{M}$ always obey $M \wedge \bar{M} = \Lambda$ and $M \vee \bar{M} = V$. Finally, we write $M_1 \leq M_2$ to indicate that all the probability in M_1 is also in M_2.

The second step is to assume there is a *measure* on M – a function $\mu: \mathsf{M} \to [0, 1]$ such that $\mu(M)$ is the measure of the probability mass M. This function must obey the following rules:

(i) $\mu(\Lambda) = 0$.

(ii) If $M \neq \Lambda$, then $\mu(M) > 0$.

(iii) $\mu(V) = 1$.

(iv) If $\{M_\gamma\}$ is disjoint, then $\mu(\vee M_\gamma) = \sum \mu(M_\gamma)$.

Notice that (iv) applies to both finite and infinite collections; the measures of disjoint probability masses always add.

Finally, we specify a mapping $\rho: 2^\Theta \to \mathsf{M}$ which satisfies three rules:

(i) $\rho(\emptyset) = \Lambda$.

(ii) $\rho(\Theta) = V$.

(iii) $\rho(A_1 \cap A_2) = \rho(A_1) \wedge \rho(A_2)$ for all $A_1, A_2 \in 2^\Theta$.

This function is called the *allocation of probability*: for each $A \in 2^\Theta$, $\rho(A)$ is the total probability mass committed to A.

The allocation $\rho: 2^\Theta \to \mathsf{M}$ does two things. It tells which probability masses are committed to which propositions, and it tells the degree of belief in each proposition. In order to tell whether a probability mass M is committed to a proposition A, we need only check whether M is included in the total probability committed to A; i.e., whether $M \leqslant \rho(A)$. In order to find the degree of belief in a proposition A, we need only find the measure of $\rho(A)$. In other words, $\mathrm{Bel}(A) = \mu(\rho(A))$ for all $A \in 2^\Theta$, or $\mathrm{Bel} = \mu \circ \rho$.

As it turns out, this formalization corresponds exactly to the structure of belief functions. In other words, whenever $\rho: 2^\Theta \to \mathsf{M}$ is an allocation of probability, the function $\mu \circ \rho$ is a belief function on 2^Θ. And any belief function on 2^Θ can be represented in this way by some complete Boolean algebra M, some measure μ on M, and some allocation $\rho: 2^\Theta \to \mathsf{M}$.

3.1.3. *Upper Probabilities*

Since the degree of belief $\mathrm{Bel}(A)$ corresponds to the degree of support for A, it is natural to think of the quantity $1 - \mathrm{Bel}(\bar{A})$ as corresponding to the degree of plausibility of A. In other words, $1 - \mathrm{Bel}(\bar{A})$ should measure the degree to which one finds A to be plausible, or the extent to which one regards A as plausible. This interpretation fits the intuitive picture of probability masses, for since $\mathrm{Bel}(\bar{A})$ measures the probability mass committed to $\bar{A}$, $1 - \mathrm{Bel}(\bar{A})$ measures the portion of our probability that is not committed to $\bar{A}$, i.e., is not committed against A.

Following A. P. Dempster, I will call the quantity $1 - \mathrm{Bel}(\bar{A})$ the upper probability of A, and denote it by $P^*(A)$. And I will call a function $P^*: 2^\Theta \to [0, 1]$ an *upper probability function* if and only if the function $\mathrm{Bel}: 2^\Theta \to [0, 1]$ defined by $\mathrm{Bel}(A) = 1 - P^*(\bar{A})$ is a belief function.

Notice that alongside the logical vocabulary of support, we now have a subjective vocabulary. The full correspondence between the two vocabularies is shown in Table IX.

TABLE IX

The two vocabularies

Subjective		Logical	
Degree of Belief in A.	$\mathrm{Bel}(A)$	Degree of Support for A.	$S(A)$
Degree of Doubt for A.	$\mathrm{Bel}(\bar{A})$	Degree of Dubiety of A.	$S(\bar{A})$
Upper Probability of A.	$1 - \mathrm{Bel}(\bar{A})$	Degree of Plausibility of A.	$1 - S(\bar{A})$

3.2. Condensability

In §1.4, I argued that a plausibility function based on empirical evidence should be condensable. The same should evidently apply to an upper probability function $P^*: 2^\Theta \to [0, 1]$ based on empirical evidence; it ought to obey

$$P^*(A) = \sup \{P^*(B) \mid B \subset A;\ B \text{ is finite}\}$$

for all $A \in 2^\Theta$.

Remarkably enough, an upper probability function P^* will obey this rule if and only if the corresponding allocation $\rho: 2^\Theta \to \mathsf{M}$ obeys

$$\rho(\bigcap A_\gamma) = \wedge\, \rho(A_\gamma)$$

for all collections $\{A_\gamma\}$ of elements of 2^Θ. In other words, the requirement of condensability corresponds to the requirement that ρ should preserve

all intersections – infinite as well as finite. (Any allocation preserves all finite intersections.)

The requirement that ρ should preserve all intersections seems to be natural and necessary if ρ is to faithfully represent the intuitive picture underlying the axioms for belief functions. For it corresponds to the intuition that a probability mass committed to each of a collection $\{A_\gamma\}$ of propositions should also be committed to their logical conjunction $\bigcap A_\gamma$. Hence condensability emerges as a condition both appropriate to empirical evidence and natural to our notion of partial belief.

3.2.1. *A Geometric Intuition*

A condensable allocation lends itself to a very vivid geometric intuition. Think of Θ as a geometric set of points, and think of our probability as being spread over the set Θ. But instead of requiring that it be distributed in a fixed way, as in the picture associated with a 'distribution of probability', let us permit it a limited freedom of movement. Indeed, each time that a probability mass is 'committed' to a set A, let us say that it is 'constrained' to A, meaning that even though it may enjoy some freedom of movement, none of it can manage to escape from A.

This picture fits our rules for the commitment of probability masses to propositions perfectly. For example, a probability mass that is constrained to stay within a subset A is obviously constrained to stay within any subset B such that $A \subset B$. And any probability mass that is constrained to stay within each of a collection $\{A_\gamma\}_{\gamma \in \Gamma}$ of subsets must stay within $\bigcap A_\gamma$.

Furthermore, the belief function and the upper probability function can be interpreted very simply in terms of this picture: the degree of belief in A is the measure of all the probability that cannot get out of A, while A's upper probability is the measure of all the probability that can get into A.

The word 'condensability' itself acquires a vivid meaning in this picture. The fact is that no matter how diffusely the probability that can get into a set A might spread itself out over A, it is always possible to 'condense' it into a (possibly countably infinite) number of discrete pieces, each of which can get into some singleton $\{\theta\}$, where $\theta \in A$.

A simple example will show how this picture can help develop our intuition for belief functions. Consider Figure 11, where a subset $A \subset \Theta$ is shown with disjoint subsets A_1 and A_2 of A such that $A_1 \cup A_2 = A$. When we think of our probability mass as being semi-mobile over Θ, we can easily see how it might happen that $\mathrm{Bel}(A_1)=0$, $\mathrm{Bel}(A_2)=0$ and yet $\mathrm{Bel}(A_1 \cup A_2)>0$. And more generally, we can see how there might often be some probability that is constrained neither to A_1 nor to A_2 yet is constrained to $A_1 \cup A_2$. Indeed, this will happen whenever a probability mass M is free to move back and forth from A_1 to A_2 but is not free to move outside of $A_1 \cup A_2 = A$.

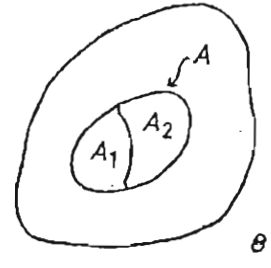

Fig. 11.

3.2.2. *Commonality Numbers*

Consider a finite collection $M_1, \ldots, M_n$ of probability masses. If one knows the measure of each of the M_i, and also how much overlap there is between each pair, among each triplet, etc., then one can deduce the measure of $M_1 \vee \ldots \vee M_n$ by the formula

$$\mu(M_1 \vee \ldots \vee M_n) = \sum_i \mu(M_i) - \sum_{i<j} \mu(M_i \wedge M_j) + \\ + - \ldots + (-1)^{n+1} \mu(M_1 \wedge \ldots \wedge M_n). \tag{22}$$

A similar formula enables one to obtain the measure of an intersection from the measures of unions:

$$\mu(M_1 \wedge \ldots \wedge M_n) = \sum_i \mu(M_i) - \sum_{i<j} \mu(M_i \vee M_j) + \\ + - \ldots + (-1)^{n+1} \mu(M_1 \vee \ldots \vee M_n). \tag{23}$$

These two relations play an important role in the theory of condensable allocations.

Returning to our geometric picture, let $\zeta(\theta)$ denote the total probability mass that can get into the singleton $\{\theta\}$, and consider the intuitive significance of unions and intersections of the probability masses $\zeta(\theta)$. The intuitive significance of a union is obvious; $\zeta(\theta_1) \vee \ldots \vee \zeta(\theta_n)$ is the total probability mass that can get into the finite set $\{\theta_1, \ldots, \theta_n\}$. The intuitive significance of an intersection is a bit more subtle, though; $\zeta(\theta_1) \wedge \ldots \wedge \zeta(\theta_n)$ is the total probability mass that can get to each and every one of the points $\theta_1, \ldots, \theta_n$ – i.e., the total probability mass that has complete freedom of movement within the set $\{\theta_1, \ldots, \theta_n\}$. It is convenient to speak of $\zeta(\theta_1) \wedge \ldots \wedge \zeta(\theta_n)$ as the probability mass that is 'common' to the points $\theta_1, \ldots, \theta_n$.

Now $\mu(\zeta(\theta))$, being the measure of the total probability that can get into $\{\theta\}$, is simply $P^*(\{\theta\})$. Similarly,

$$P^*(\{\theta_1, \ldots, \theta_n\}) = \mu(\zeta(\theta_1) \vee \ldots \vee \zeta(\theta_n)).$$

The quantities $\mu(\zeta(\theta_1) \wedge \ldots \wedge \zeta(\theta_n))$, on the other hand, are new to us. I will denote

$$Q(\{\theta_1, \ldots, \theta_n\}) = \mu(\zeta(\theta_1) \wedge \ldots \wedge \zeta(\theta_n)),$$

and I will call $Q(\{\theta_1, \ldots, \theta_n\})$ the *commonality number* for $\{\theta_1, \ldots, \theta_n\}$.

The relations (22) and (23) provide us, of course, with the connections between upper probabilities and commonality numbers:

$$P^*(\{\theta_1, \ldots, \theta_n\}) = \sum_i Q(\{\theta_i\}) - \sum_{i<j} Q(\{\theta_i, \theta_j\}) + - \ldots + + (-1)^{n+1} Q(\{\theta_1, \ldots, \theta_n\}),$$

and

$$Q(\{\theta_1, \ldots, \theta_n\}) = \sum_i P^*(\{\theta_i\}) - \sum_{i<j} P^*(\{\theta_i, \theta_j\}) + - \ldots + + (-1)^{n+1} P^*(\{\theta_1, \ldots, \theta_n\}).$$

Or, in a notation that is sometimes more convenient:

$$(24) \qquad P^*(A) = \sum_{\substack{T \subset A \\ T \neq 0}} (-1)^{1+\operatorname{card} T} Q(T)$$

and

$$Q(A) = \sum_{\substack{T \subset A \\ T \neq 0}} (-1)^{1+\operatorname{card} T} P^*(T)$$

for all finite subsets A of Θ.

Formula (24) means in particular that the upper probabilities of finite subsets are determined by the commonality numbers. In the condensable case this implies that the entire upper probability function is determined by the commonality numbers. In fact a condensable upper probability function will obey

$$P^*(A) = \sup_{\substack{B \subset A \\ B \text{ finite}}} P^*(B),$$

or

$$P^*(A) = \sup_{\substack{B \subset A \\ B \text{ finite}}} \sum_{\substack{T \subset B \\ T \neq 0}} (-1)^{1+\operatorname{card} T} Q(T)$$

for all $A \in 2^\Theta$.

So commonality numbers provide us with yet another way of specifying a condensable belief function. As it turns out, they provide the easiest way to specify many important condensable belief functions. And they also provide the simplest way of expressing Dempster's rule of combination.

3.3. Dempster's rule of combination

In §1.2, I illustrated Lambert's rule for combining support functions over two alternatives, and I mentioned a more general rule, due to A. P. Dempster, which applies when Θ is larger provided the support functions satisfy certain auxiliary conditions. As it turns out, these auxiliary conditions are precisely the rules for condensable belief functions.

The felicitousness of these rules is hardly surprising. For the crucial step in applying Lambert's rule is the representation of each support function by an imaginary mass, some of which is committed to each alternative. And as we have seen, the rules for condensable belief functions are precisely the rules that make an effective representation by imaginary 'probability masses' possible.

The actual derivation of Dempster's rule is rather complicated. After establishing the existence of a 'Boolean algebra of probability masses' representing each belief function, one must 'orthogonally combine' the two Boolean algebras of probability masses in some way analogous to the orthogonal combination of the two line segments in §1.2. One must then determine which of the resulting probability masses are committed to which elements of 2^Θ, and which are committed contradictorily and hence cannot be counted.

Let me describe the result that finally emerges from this process.

Suppose one combines the two condensable belief functions Bel_1: $2^\Theta \to [0, 1]$ and Bel_2: $2^\Theta \to [0, 1]$. Then the result of the combination is a condensable belief function Bel: $2^\Theta \to [0, 1]$ given by

$$\mathrm{Bel}(A) = \frac{c(A) - c(\emptyset)}{1 - c(\emptyset)}, \tag{25}$$

where

$$c(A) = \sup\Big\{\sum_i \mathrm{Bel}_1(A_i)\,\mathrm{Bel}_2(B_i) - \sum_{i<j} \mathrm{Bel}_1(A_i \cap A_j)\,\mathrm{Bel}_2(B_i \cap B_j) + - \ldots + (-1)^{n+1}\,\mathrm{Bel}_1(A_1 \cap \ldots \cap A_n)\,\mathrm{Bel}_2(B_1 \cap \ldots \cap B_n)\Big\},$$

the supremum being taken over all collections $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ of elements of 2^Θ such that $A_i \cap B_i \subset A$ for each i.

The only case in which two condensable belief functions Bel_1 and Bel_2 cannot be combined is when they flatly contradict each other – i.e., when there exists $A \in 2^\Theta$ such that $\mathrm{Bel}_1(A) = 1$ and $\mathrm{Bel}_2(\bar{A}) = 1$. When such a contradiction occurs, $c(\emptyset) = 1$, and (25) cannot be applied. As long as such a contradiction does not occur, however, $c(\emptyset) < 1$ and (25) can be applied.

The derivation of the commonality numbers for Bel from the commonality numbers for Bel_1 and Bel_2 is quite simple; except for a constant of renormalization, one simply multiplies. More precisely, if the commonality numbers for Bel_1, Bel_2 and Bel are denoted by $Q_1(A)$, $Q_2(A)$ and $Q(A)$, respectively, then

$$Q(A) = k\,Q_1(A)\,Q_2(A)$$

for all $A \in 2^\Theta$, where

$$k = \frac{1}{1 - c(\emptyset)}.$$

The constant k is also determined by the requirement that

$$P^*(\Theta) = \sup_{\substack{A \subset \Theta \\ A \text{ finite}}} \sum_{\substack{T \subset A \\ T \neq \emptyset}} (-1)^{1 + \mathrm{card}\,T} Q(T) = k \sup_{\substack{A \subset \Theta \\ A \text{ finite}}} \sum_{\substack{T \subset A \\ T \neq \emptyset}} (-1)^{1 + \mathrm{card}\,T} Q_1(T)\,Q_2(T) = 1.$$

The actual computation of k is often quite difficult.

Although I have presented (25) as a rule for combining condensable belief functions, it can sometimes be applied in the noncondensable case. Unfortunately, its general usefulness in the noncondensable case is questionable. It cannot always be applied, for the lack of condensability will sometimes allow $c(\emptyset) = 1$ even when there is no apparent contradiction. And even when it can be applied, the interpretation of the results may be problematic.

3.3.1. *The Murder of Mr Green*

A simple example will do more to convey the nature of Dempster's rule than the preceding formulae.

Mr Green has been murdered, and we are certain that the murder was committed by either Dr Black, Dr Gray, or Mr White. Besides the evidence that allows us to narrow the field to these three suspects, we have evidence based on the mode of the murder, and evidence based on motives for the murder.

Mr Green was poisoned by a rare chemical. This fact provides Mr White's strongest defense, for there seems to have been little way he could have obtained the chemical. But as physicians, both Dr Black and Dr Gray had access to the chemical. Of these two, Dr Black is particularly implicated, for chemicals seem to be tightly controlled at the hospital where Dr Gray works.

When we consider motives, we obtain quite a different picture. There is no apparent motive that can be ascribed to Dr Black. On the other hand, both Dr Gray and Mr White are alleged to have been Mrs Green's lovers, and hence might have plotted with Mrs Green to inherit Mr Green's fortune. The evidence is weaker in the case of Dr Gray, but Mrs Green's involvement with Mr White was practically public knowledge.

In order to apply our theory to this example, we must set $\Theta = \{\text{Black, Gray, White}\}$ and postulate support functions S_1 and S_2 based on the two separate sources of evidence.

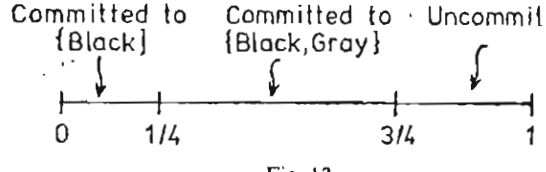

Fig. 12.

The support function S_1, derived from the mode of the murder, is represented in Figure 12. According to this figure, 3/4 of our probability is committed to the guilt of one of the doctors, and part of that is committed more specifically to the guilt of Dr Black. Explicitly,

$$S_1(\{\text{Black}\}) = 1/4 \qquad S_1(\{\text{Black, Gray}\}) = 3/4$$
$$S_1(\{\text{Gray}\}) = 0 \qquad S_1(\{\text{Black, White}\}) = 1/4$$
$$S_1(\{\text{White}\}) = 0 \qquad S_1(\{\text{Gray, White}\}) = 0.$$

And hence

$$PL_1(\{\text{Gray, White}\}) = 3/4 \qquad PL_1(\{\text{White}\}) = 1/4$$
$$PL_1(\{\text{Black, White}\}) = 1 \qquad PL_1(\{\text{Gray}\}) = 3/4$$
$$PL_1(\{\text{Black, Gray}\}) = 1 \qquad PL_1(\{\text{Black}\}) = 1$$

The support function S_2, derived from the consideration of motives, is similarly represented in Figure 13. In view of their possible motives, we have commited 2/3 of our probability to the proposition that either

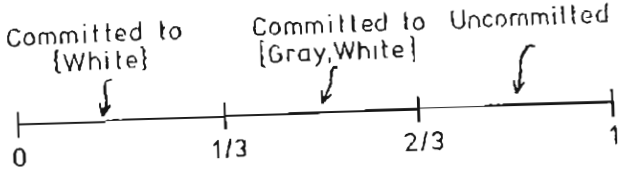

Fig. 13.

Dr Gray or Mr White is the murderer. One-half of this, or 1/3 of our probability, is committed more specifically to Mr White, whose involvement with Mrs Green is certain. Explicitly,

$$S_2(\{\text{Black}\}) = 0 \qquad S_2(\{\text{Black, Gray}\}) = 0$$
$$S_2(\{\text{Gray}\}) = 0 \qquad S_2(\{\text{Black, White}\}) = 1/3$$
$$S_2(\{\text{White}\}) = 1/3 \qquad S_2(\{\text{Gray, White}\}) = 2/3.$$

And hence

$$PL_2(\{\text{Gray, White}\}) = 1 \qquad PL_2(\{\text{White}\}) = 1$$
$$PL_2(\{\text{Black, White}\}) = 1 \qquad PL_2(\{\text{Gray}\}) = 2/3$$
$$PL_2(\{\text{Black, Gray}\}) = 2/3 \qquad PL_2(\{\text{Black}\}) = 1/3.$$

The combination of S_1 and S_2, as illustrated in Figure 14, is quite analogous to the application of Lambert's rule in §1.2. The only novelty

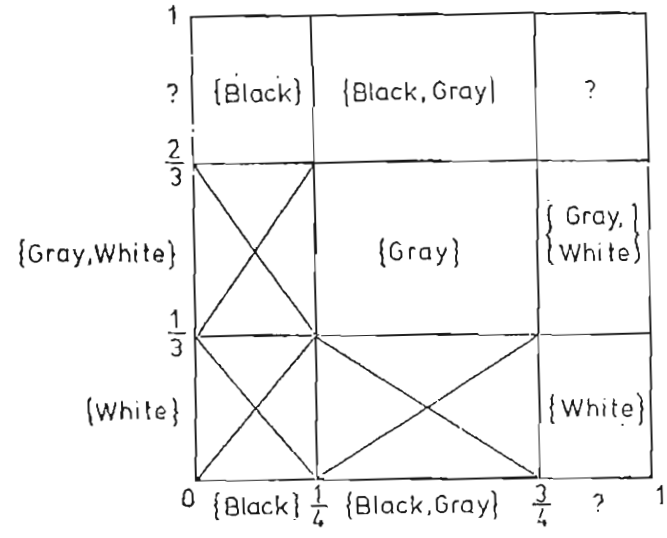

Fig. 14.

occurs in assigning the central rectangle in Figure 14. S_1 commits that rectangle to the proposition that either Black or Gray is the murderer, while S_2 commits it to the proposition that either Gray or White is the murderer; so together they commit it to the conjunction of these two propositions – the proposition that Gray is the murderer.

The support function S based on this combined evidence is thus as follows

$$S(\text{Black}) = 1/8 \qquad S(\text{Black, Gray}) = 5/8$$
$$S(\text{Gray}) = 1/4 \qquad S(\text{Black, White}) = 1/4$$
$$S(\text{White}) = 1/8 \qquad S(\text{Gray, White}) = 1/2.$$

And the corresponding plausibility function is given by

$$PL(\text{Gray, White}) = 7/8 \qquad PL(\text{White}) = 3/8$$
$$PL(\text{Black, White}) = 3/4 \qquad PL(\text{Gray}) = 3/4$$
$$PL(\text{Black, Gray}) = 7/8 \qquad PL(\text{Black}) = 1/2.$$

Notice that Dr Gray is the most seriously implicated suspect on the basis of the combined evidence, even though neither of the separate sources of evidence gave positive support to his guilt. It is the combination of means and motive that points strongly in his direction.

Notice that both PL_1 and PL_2 are consonant, while PL is dissonant.

3.3.2. *The Combination of Evidence*

Dempster's rule of combination is best understood as a rule for combining separate bodies of evidence. For reasons that are now evident, I have presented it as a rule for combining 'belief functions', but I do not want to suggest that the rule provides a method for 'pooling beliefs'. Evidence can be pooled; beliefs cannot be.

If two individuals' belief functions are based on separate sources of evidence, then those individuals can pool their evidence by using Dempster's rule to combine their belief functions. But if the two belief functions are based even partially on the same evidence, then Dempster's rule is inappropriate and will give misleading results. For it always treats the separate belief functions it combines as if they were based on separate sources of evidence.

Suppose, for example, that $\Theta=\{\theta_1, \theta_2\}$ and that the two individuals have exactly the same belief function, namely $\mathrm{Bel}_0: 2^\Theta \to [0, 1]$, where $\mathrm{Bel}_0(\{\theta_1\})=2/10$ and $\mathrm{Bel}_0(\{\theta_2\})=1/10$. Now if we combine Bel_0 with itself we will obtain a belief function that indicates, roughly speaking, twice as much evidence in both directions as Bel_0 indicates. Indeed, the belief function Bel obtained by combining Bel_0 with itself is given by $\mathrm{Bel}(\{\theta_1\})=\frac{3}{9}\frac{2}{6}$ and $\mathrm{Bel}(\{\theta_2\})=\frac{1}{9}\frac{5}{6}$. Clearly, the two individuals should retain the belief function Bel_0 rather than adopt Bel.

3.4. Conditioning belief functions

Suppose we begin with a belief function $\mathrm{Bel}: 2^\Theta \to [0, 1]$ and then obtain new evidence showing that the true value of θ is not only in Θ but also in some proper subset Θ_0 of Θ. Can we use Dempster's rule of combination to combine this new knowledge with the evidence underlying Bel?

We can, provided that we can represent the new evidence by a belief function. But that can be done quite simply; the knowledge that the true value of θ is in Θ_0 is conveyed by the belief function $\mathrm{Bel}_0: 2^\Theta \to [0, 1]$, where

$$\mathrm{Bel}_0(A) = \begin{cases} 1 & \text{if } \Theta_0 \subset A \\ 0 & \text{otherwise} \end{cases}.$$

So we should use Dempster's rule to combine Bel_0 and Bel.

This combination results in the belief function $\mathrm{Bel}(\cdot \mid \Theta)$, given by

$$\text{(26)} \qquad \mathrm{Bel}(A \mid \Theta_0) = \frac{\mathrm{Bel}(A \cup \bar{\Theta}_0) - \mathrm{Bel}(\bar{\Theta}_0)}{1 - \mathrm{Bel}(\bar{\Theta}_0)}$$

for all $A \in 2^\Theta$. A more succinct formula can be given in terms of the upper probability functions. If we denote the upper probability functions corresponding to Bel and $\mathrm{Bel}(\cdot \mid \Theta_0)$ by P^* and $P^*(\cdot \mid \Theta_0)$, respectively, then (26) becomes

$$P^*(A \mid \Theta_0) = \frac{P^*(A \cap \Theta_0)}{P^*(\Theta_0)}$$

for all $A \in 2^\Theta$.

This rule for obtaining $\mathrm{Bel}(\cdot \mid \Theta_0)$ from Bel is called *Dempster's rule of conditioning*, and the symbols '$\mathrm{Bel}(A \mid \Theta_0)$' may be read 'the degree of belief in A, given Θ_0'.

Since this rule is a special case of Dempster's rule of combination, it can also be expressed in terms of the commonality members. Indeed, if the commonality numbers for Bel are denoted by $Q(A)$ and those for $\mathrm{Bel}(\cdot \mid \Theta_0)$ are denoted by $Q^0(A)$, then $Q(A)$ and $Q^0(A)$ will be related by

$$Q^0(A) = \begin{cases} \dfrac{Q(A)}{P^*(\Theta_0)} & \text{if } A \subset \Theta_0 \\ 0 & \text{otherwise} \end{cases}.$$

3.4.1. *The Analogy with Bayes' Theory*

The general notion of conditioning, and formula (27) in particular, is strongly reminiscent of Bayes' theory. That theory explores the possibility of expressing degrees of belief or support by functions $P: 2^\theta \to [0, 1]$ that obey the rules

(i) $P(\emptyset) = 0$,

(ii) $P(\Theta) = 1$,

(iii) $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$;

and it has at its core a rule of conditioning closely analogous to (27). By that rule, conditioning P on Θ_0 will produce another function obeying (i)-(iii) and given by the formula

$$\text{(28)} \qquad P(A \mid \Theta_0) = \frac{P(A \cap \Theta_0)}{P(\Theta_0)}$$

for all $A \in 2^\Theta$.

Actually, Bayes' theory can be regarded as a special case of our theory of belief, and (28) can be regarded as the corresponding special case of (26). For any function satisfying (i)–(iii) will qualify as a belief function, and in the case of a belief function P obeying (iii), (26) reduces to (28). Actually, a belief function obeying (iii) is identical with its upper probability function, so both (26) and (27) reduce to (28).

Students of Bayes' theory have explained (28) in various ways, but one of the most appealing explanations is in terms of probability masses. In these terms, rule (iii) corresponds to the situation where our probability, instead of being allowed some freedom of movement, is distributed in a fixed way over Θ. In such a situation, conditioning on Θ_0 can be thought of as discarding the probability that is distributed over $\bar{\Theta}_0$ and 'renormalizing' the measure of the rest.

We can verify that such a procedure produces (28) by studying Figure 15. Referring to that figure, we see that when the probability distributed over $\bar{\Theta}_0$ is discarded, part of that which was distributed over A will be

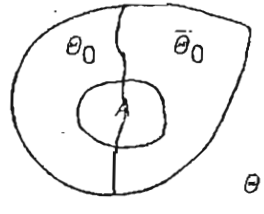

Fig. 15.

included in what is discarded, so that only the probability that was distributed over $A \cap \Theta_0$ will remain. This accounts for the numerator in (28). The denominator results from the 'renormalization'. The difficulty is that when the probability mass over $\bar{\Theta}_0$ is discarded, the total remainder will have a measure of only $1-P(\bar{\Theta}_0)=P(\Theta_0)$. Since we want our probability to have total measure 1, we must multiply the measures of all our probability masses by $(P(\Theta_0))^{-1}$.

3.4.2. *The Geometric Argument for* (26)

This intuitive argument can be generalized to explain our rule of conditioning. Indeed, when we condition a belief function Bel: $2^\Theta \to [0, 1]$

on Θ_0, how must we treat our semi-mobile probability masses? Clearly, we should eliminate the probability that is committed to Θ_0 and renormalize the measure of the remainder. The measure of the probability thus eliminated will be $\mathrm{Bel}(\bar{\Theta}_0)$, so the constant of renormalization will be $(1-\mathrm{Bel}(\bar{\Theta}_0))^{-1}$. There is only one further idea that must be introduced: since our probability is allocated in a semimobile way over Θ rather than being distributed in a fixed way, we must recognize that the restriction to Θ_0 may further restrict the mobility of some of our probability without eliminating it entirely. This means that some of our probability that was not committed to A before may become committed to A by the restriction to Θ_0.

In fact, any probability that was committed to $A \cup \bar{\Theta}_0$ before will now be committed to A, unless it is eliminated. In general, then, the amount of probability committed to A after conditioning will be the measure of the probability previously committed to $A \cup \bar{\Theta}_0$ less the measure of the probability eliminated, or

$$\mathrm{Bel}(A \cup \bar{\Theta}_0) - \mathrm{Bel}(\bar{\Theta}_0).$$

But of course this must be renormalized, so we obtain

$$\mathrm{Bel}(A \mid \Theta_0) = \frac{\mathrm{Bel}(A \cup \bar{\Theta}_0) - \mathrm{Bel}(\bar{\Theta}_0)}{1 - \mathrm{Bel}(\bar{\Theta}_0)}$$

for all $A \in 2^\Theta$.

4. Statistical Support

Our present goal is to define statistical support functions that obey both the postulates of plausibility and the rules for condensable belief functions. In this section, I will present two different methods for doing so. These two methods are based on quite different rationales, and produce different systems of support functions.

The first method is based on the assumption that support functions from single observations should be consonant. This assumption is contrary to the opinions I expressed in §2, but it does not contradict the postulates of plausibility. It leads to the *linear support functions*.

The second method is based on a theory of partial belief that embraces aleatory laws. It leads to the *simplicial support functions*, which are both more interesting and more difficult to study than the linear ones.

In §2, I distinguished between elements of the set Θ and the aleatory laws $\{P_\theta\}_{\theta\in\Theta}$ corresponding to them. Such a distinction serves to emphasize that the true value of θ may have a broader substantive significance than the aleatory law P_θ that it associates with the particular experiment at hand. In particular, it serves to remind us that two distinct values θ_1 and θ_2 might have $P_{\theta_1} \equiv P_{\theta_2}$. Unfortunately, though, the distinction complicates our notation. In this section, I will dispense with the distinction and think of each element θ of Θ as an aleatory law $\theta: X \to [0, 1]$.

In this notation, the postulates of plausibility read as follows:

Suppose Θ is the complete statistical specification on the discrete space X, and suppose $PL_x: 2^\Theta \to [0, 1]$ is the plausibility function based on the single observation $x \in X$. Then

(I) $PL_x(\{\theta\}) = c(x)\,\theta(x)$, where $c(x)$ depends on x but not on θ. And

(II) $PL_x(\{\theta_1, \theta_2\}) = PL_x(\{\theta_1\})$ whenever $\dfrac{\theta_1(x)}{\theta_1(y)} \geqslant \dfrac{\theta_2(x)}{\theta_2(y)}$ for all y.

And the stronger versions of these postulates become:

Suppose Θ is a statistical specification on the discrete space X, and suppose $PL_{\mathbf{x}}: 2^\Theta \to [0, 1]$ is the plausibility function based on the observations $\mathbf{x} = (x_1, \ldots, x_n)$. Then

(I′) $PL_{\mathbf{x}}(\{\theta\}) = c(\mathbf{x})\,\theta(x_1) \ldots \theta(x_n)$, where $c(\mathbf{x})$ depends on $\mathbf{x}$ but not on θ. And

(II′) $PL_{\mathbf{x}}(\{\theta_1, \theta_2\}) = PL_{\mathbf{x}}(\{\theta_1\})$ whenever $\dfrac{\theta_1(x_1) \ldots \theta_1(x_n)}{\theta_1(y_1) \ldots \theta_1(y_n)} \geqslant \dfrac{\theta_2(x_1) \ldots \theta_2(x_n)}{\theta_2(y_1) \ldots \theta_2(y_n)}$ for all y_i from X.

4.1. The three postulates of support

When I say that our statistical support functions should obey the rules for condensable belief functions, I mean both that they should qualify as condensable belief functions and that they should obey Dempster's rules of combination and conditioning.

The relevance of Dempster's rule of combination is obvious. If $\mathbf{x} = (x_1, \ldots, x_n)$, then the support function $S_{\mathbf{x}}$ resulting from the n observations ought to be the same as the support function obtained by combining the n support functions $S_{x_1}, \ldots, S_{x_n}$ by Dempster's rule of combination.

Dempster's rule of conditioning is relevant because some statistical specifications are subsets of others. Suppose, indeed, that Θ is the complete statistical specification on X and we obtain a support function $S_{\mathbf{x}}: 2^\Theta \to [0, 1]$ based on the observations $\mathbf{x}$. And suppose that we then change our minds and want to consider the restricted specification $\Theta_0 \subset \Theta$. Then according to our theory of partial belief, we should obtain our support function on 2^{Θ_0} by conditioning $S_{\mathbf{x}}$ on Θ_0.

Notice that Dempster's rules reduce our problem to that of finding support functions in the special case of a single observation from a complete specification. It is for this reason that I have stated our two postulates of plausibility, (I) and (II), in terms of this special case.

So in addition to (I) and (II), we now have three postulates of support:

(III) *The First Postulate of Support.* Suppose Θ is the complete statistical specification on a discrete space X. Then the support function $S_X: 2^\Theta \to [0, 1]$ based on an observation $x \in X$ must be a condensable belief function.

(IV) *The Second Postulate of Support.* Suppose Θ is the complete statistical specification on a discrete space X, and $S_{x_1}, \ldots, S_{x_n}$ are the support functions on 2^Θ based on the observations $x_1, \ldots, x_n$, respectively. Then the support function $S_X: 2^\Theta \to [0, 1]$ based on all the observations $\mathbf{x} = (x_1, \ldots, x_n)$ must be the belief function obtained by combining $S_{x_1}, \ldots, S_{x_n}$ by Dempster's rule of combination.

(V) *The Third Postulate of Support.* Suppose Θ is the complete statistical specification on a discrete space X, and $S_{\mathbf{x}}: 2^\Theta \to [0, 1]$ is the support function based on the observations $\mathbf{x} = (x_1, \ldots, x_n)$. Then the support function $S_{\mathbf{x}}^{\Theta_0}: 2^{\Theta_0} \to [0, 1]$ based on the observations $\mathbf{x}$ and the restricted specification $\Theta_0 \subset \Theta$ must be the belief function obtained by conditioning $S_{\mathbf{x}}$ on Θ_0 by Dempster's rule of conditioning.

These are the further postulates that I promised in §2; they allow one to deduce (I′) from (I) and (II′) from (II).

They also allow one to strengthen (II) so as to apply to larger subsets than doubletons:

(II″). Suppose $A \subset B \subset \Theta$, and suppose that for every $\theta_2 \in B$ there exists $\theta_1 \in A$ such that

$$\frac{\theta_1(x)}{\theta_1(y)} \geqslant \frac{\theta_2(x)}{\theta_2(y)}$$

for all $y \in X$. Then $PL_x(B) = PL_x(A)$.

In other words, the enlargement of a hypothesis cannot increase its plausibility as long as for each now simple hypothesis added there is already a simple hypothesis present under which the actual observation has greater likelihood relative to every other possible observation.

4.1.1. *The Dissonance of Statistical Evidence*

In general, our five postulates do not suffice to completely determine the support functions S_x and S_x. They are obeyed, for example, by both the linear and the simplicial support functions, and as we will see below, these differ in important respects. But any method of computing support functions that obeys the five postulates will emphasize the dissonant nature of statistical evidence.

It is Dempster's rule of combination that is responsible for this prominence of dissonance. As more and more observations are accumulated, this rule will operate to produce more and more dissonance. And eventually it will lead to a support function indicating strong evidence against each possibility – i.e., against each possible aleatory law. In other words, the plausibility of every simple hypothesis will decline; typically, the plausibility of the most plausible simple hypothesis will tend to zero as the number of observations grows.

In this respect, the present theory contrasts sharply with some other methods of assessing statistical evidence, which seem to treat it as consonant. The method of nested confidence regions provides a case in point. The theory of confidence regions is, of course, an operational rather than an epistemic theory. But nested confidence regions corresponding to different confidence coefficients are often used informally to summarize statistical evidence. (Cf. p. 62 of Erich Lehman's book.) Such a nested family can sometimes be interpreted as defining a support function: the degree of support for a given subset A of Θ will be the largest confidence coefficient such that A contains the confidence region with that coefficient. A support function so defined will be consonant, no matter how conflicting the evidence may appear to be.

4.1.2. *The Binomial Specification*

While our five postulates do not uniquely determine the support functions S_x and S_x in general, they do uniquely determine them in the case where X has only two elements.

Suppose, indeed, that X has two elements and that Θ is the complete specification on X. Then it may be deduced from (I) and (II″) that PL_x must be given by

$$PL_x(A) = \sup_{\theta \in A} \theta(x)$$

for all $A \in 2^\Theta$. The plausibility functions for restricted specifications and/or many observations can be obtained, of course, by combination and conditioning; they are the upper probability functions for which A. P. Dempster gave detailed formulae in his 1966 article.

4.2. The linear plausibility functions

In §2.1, I argued that statistical evidence can be dissonant even when it consists of only a single observation. It is possible, however, to take the opposite view and insist that plausibility functions based on a single observation should always be consonant. It turns out that this assumption of consonance suffices to completely determine the plausibility function based on a single observation $x \in X$ and a complete specification Θ on X. Indeed, that plausibility function will be given by

$$PL_x(A) = \sup_{\theta \in A} \theta(x) \tag{29}$$

for all $A \in 2^\Theta$, just as in the special case of the binomial specification.

In the case of a restricted specification, Θ_0, we must condition (29), obtaining

$$PL_x(A) = \frac{\sup_{\theta \in A} \theta(x)}{\sup_{\theta \in \Theta_0} \theta(x)} \tag{30}$$

for all $A \in 2^{\Theta_0}$. This too will be a consonant plausibility function.

Using Dempster's rule of combination, we can also obtain the plausibility functions based on many observations. Hence the assumption of consonance for single observations, together with our five postulates, determines a complete system of plausibility functions. I call them the *linear plausibility functions*.

The simplicity of (29) and (30) is a strong point in favor of the linear plausibility functions. The linear plausibility functions for multiple observations are somewhat more complicated, but still manageable. This simplicity contrasts with the complexity of the simplicial plausibility functions, described in §4.3 below. The simplicial plausibility functions are theoretically more attractive than the linear ones, but they present formidable computational difficulties.

4.2.1. *The Continuous Case*

Thus far I have dealt exclusively with aleatory laws on discrete spaces. It is possible, however, to define the linear plausibility functions for the continuous case as well.

A continuous statistical specification on a space X is usually described with reference to a fixed measure on X; each aleatory law is given by a density with respect to that measure. In the simplest case, X is the real line and the reference measure is Lebesque measure. In that case, an aleatory law given by a non-negative function θ such that

$$\int_{-\infty}^{\infty} \theta(x)\,dx = 1\,;$$

θ is called a *probability density*, and it is understood to assign probability

$$\int_{A} \theta(x)\,dx$$

to any (measurable) subset A of the real line.

This description makes no reference to any topology on X. On the other hand, it is commonly and quite correctly argued that continuous statistical specifications are extremely idealized and should be understood as limiting cases of discrete specifications. And the notion of approximating a continuous specification by a discrete one can be made intelligible only in the context of a topology on X and some requirements of continuity on the densities in the specification.

It is not clear just what requirements of continuity should be imposed, but in the case of the real line, we might require each density to be continuous, with the moduli of continuity for the different densities bounded at each point. More precisely, we might say that a set Θ_0 of functions on the real line X is a statistical specification on X provided that

(i) $\theta(x) \geqslant 0$ for all $x \in X$ and $\theta \in \Theta_0$.

(ii) $\int_{-\infty}^{\infty} \theta(x) = 1$ for all $\theta \in \Theta_0$.

(iii) For each $x \in X$ there exists $K < \infty$ such that
$|\theta(x') - \theta(x)| < K|x - x'|$ for all $x' \in X$ and $\theta \in \Theta_0$.

One consequence of these requirements is that $\sup_{\theta \in \Theta_0} \theta(x) < \infty$ for each $x \in X$. So whenever these requirements are met, (30) can be used to define the plausibility function $PL_x\colon 2^{\Theta_0} \to [0, 1]$ based on the observation $x \in X$. And plausibility functions based one many observations can then be obtained by combination.

4.3. The simplicial plausibility functions

As I mentioned above, the simplicial support functions can be justified by a theory of belief that embraces aleatory laws. In this section, I will sketch this theory of belief and then briefly describe the simplicial plausibility functions.

4.3.1. *Aleatory Laws and Degrees of Belief*

The idea of an aleatory law has been intertwined with the idea of degrees of belief throughout the history of both ideas, but the connection can be perplexing. For though an aleatory law can always supply us with degrees of belief, it is appropriate to adopt those degrees of belief only when we know the law to hold. And such knowledge is rarely available.

Let me be more precise. If an experiment has outcomes in X, and we know it is governed by the aleatory law $\theta\colon X \to [0, 1]$, then we will naturally adopt $\theta(x)$ as our degree of belief that the experiment will

result in $x \in X$. And more generally, we will adopt $\sum_{x \in A} \theta(x)$ as our degree of belief that the experiment will result in one of possible outcomes in a subset A of X. In other words, we will adopt the belief function $\mathrm{Bel}_\theta: 2^X \to [0, 1]$ given by $\mathrm{Bel}_\theta(A) = \sum_{x \in A} \theta(x)$. In fact, though, we can never really be certain that the experiment is governed by the aleatory law θ. So the belief function Bel_θ is appropriate only conditionally upon knowledge that we do not and cannot have.

This perplexing situation could be made intelligible within our theory of belief if we could somehow represent our knowledge by another belief function which produces Bel_θ only when conditioned on the fact that θ is the true aleatory law governing the experiment. Is this possible?

The belief function Bel_θ applies to propositions about the outcome of a forthcoming trial, and we wish to obtain it by conditioning another belief function, say Bel, on the proposition that the aleatory law θ governs the experiment. Hence Bel must apply both to propositions about the outcome of the trial and to propositions about what aleatory law governs the experiment. And it must also apply to propositions that simultaneously assert something about the outcome of the trial and something about which aleatory law governs the experiment.

We are led, then, to postulate the existence of a belief function $\mathrm{Bel}: 2^{\Theta \times X} \to [0, 1]$, where Θ is the collection of all possible aleatory laws on X, $\Theta \times X$ is the Cartesian product of Θ and X, and $A \in 2^{\Theta \times X}$ is taken to be the proposition that the pair (θ, x) is in $A \subset \Theta \times X$, where θ is the true aleatory law governing the experiment and x is the outcome of the forthcoming trial. Such a belief function Bel can indeed be conditioned on the proposition that a given aleatory law $\theta \in \Theta$ is the true one; we simply condition Bel on the subset $\{\theta\} \times X$ of $\Theta \times X$, obtaining a belief function on $2^{\{\theta\} \times X}$. A belief function on $2^{\{\theta\} \times X}$ amounts to the same thing as a belief function on 2^X, so we can require this belief function to be Bel_θ, as given above.

The notion of such an overall belief function Bel is very fruitful, for we can express many of our other ideas in terms of Bel. Lack of any prior opinion about the identity of the true aleatory law can be expressed, for example, by saying that $P^*(\{\theta\} \times X) = 1$ for all $\theta \in \Theta$. And most importantly, it becomes natural to obtain our support function S_x based on the observation x by conditioning Bel on x – i.e., on $\Theta \times \{x\}$. Hence conditions on the support functions S_x become conditions on Bel.

So for every discrete space X, we find ourselves demanding a function $\mathrm{Bel}: 2^{\Theta \times X} \to [0, 1]$, where Θ is the set of all aleatory laws on X, and Bel satisfies at least the following four conditions:

(1) Bel is a condensable belief function.

(2) $P^*(\{\theta\} \times X) = 1$ for all $\theta \in \Theta$, where P^* is the upper probability function corresponding to Bel.

(3) For every $\theta \in \Theta$, conditioning Bel on θ results in the belief function $\mathrm{Bel}_\theta: 2^X \to [0, 1]$ given by $\mathrm{Bel}_\theta = \sum_{x \in A} \theta(x)$ for all $A \in 2^X$.

(4) For every $x \in X$, conditioning Bel on x results in an upper probability function on 2^Θ that satisfies the second postulate of plausibility for the observation x. (More precisely, if $PL_x: 2^\Theta \to [0, 1]$ is the upper probability function obtained by conditioning Bel on $\{x\} \times \Theta$, then PL_x should satisfy $PL_x(\{\theta_1, \theta_2\}) = PL_x(\{\theta_1\})$ for all doubletons $\{\theta_1, \theta_2\} \in 2^\Theta$ such that $\theta_1(x)/\theta_1(y) \geq \theta_2(x)/\theta_2(y)$ for all $y \in X$.)

It turns out that for every discrete space X there is one and only one belief function $\mathrm{Bel}: 2^{\Theta \times X} \to [0, 1]$ satisfying these four requirements. I will not prove this fact here, but I will describe the upper probability functions $\{PL_x\}_{x \in X}$ that result from conditioning this unique belief function Bel on the various possible observations $x \in X$. These upper probability functions PL_x are, of course, the *simplicial plausibility functions*.

4.3.2. *Some Formulae*

Let me begin my description of the functions PL_x by supplying some formulae. I will then turn to a more illuminating geometric description.

I continue to denote by Θ the set of all aleatory laws on the discrete space X. Fix $x \in X$, and for each finite non-empty subset A of Θ, set

$$(31) \qquad Q_x(A) = \begin{cases} \dfrac{1}{\displaystyle\sum_{y \in X} \max_{\theta \in A} \dfrac{\theta(y)}{\theta(x)}} & \text{if } \theta(x) > 0 \text{ for all } \theta \in A, \\ 0 & \text{otherwise.} \end{cases}$$

ıese are the commonality numbers for the simplicial plausibility func-
n PL_x. In other words,

$$PL_x(A) = \sum_{\substack{T \subset A \\ T \neq 0}} (-1)^{1+\operatorname{card} T} Q_x(T)$$

r all finite non-empty subsets A of Θ. And, of course, the values of $L_x(A)$ for infinite subsets A are given by condensability:

$$PL_x(A) = \sup_{\substack{A' \subset A \\ A' \subset \text{finite}}} PL_x(A').$$

The commonality numbers $Q_x^0(A)$ and the plausibility function PL_x^0 ɔr a restricted specification $\Theta_0 \subset \Theta$ can be obtained from these formulae y the rule of conditioning. The commonality numbers $Q_x^0(A)$ for finite ıbsets A of Θ_0 will be given, for example, by

$$(32) \qquad Q_x^0(A) = \begin{cases} \displaystyle\sum_{y \in X} \max_{\theta \in A} \frac{0(y)}{0(x)}^{k} & \text{if} \quad 0(x) > 0 \quad \text{for all} \quad \theta \in A, \\ 0 & \text{otherwise,} \end{cases}$$

vhere the constant k is equal to $(PL_x(\Theta_0))^{-1}$.

1.3.3. *The Geometric Representation*

n the case where X has three elements, the function PL_x that we have just lefined can be described in terms of the geometric picture developed in §2.3.

Suppose, indeed, that $X = \{1, 2, 3\}$. Then as we saw in §2.3, the aleatory laws on X are in a one-to-one correspondence with the points of an equilateral triangle of unit area. Let me review that one-to-one correspondence, using a slightly different notation.

First, denote by M the set of measurable subsets of the triangle, and by $\mu(M)$ the measure of a subset $M \in \mathsf{M}$. Then note that the point θ of the triangle determines three smaller triangles α_θ, β_θ and γ_θ, as in Figure 16. The barycentric coordinates of θ are then

$$(\mu(\alpha_\theta), \mu(\beta_\theta), \mu(\gamma_\theta));$$

these three numbers are always non-negative and add to one. Finally, the aleatory law $\theta: X \to [0, 1]$ corresponding to the point θ of the triangle

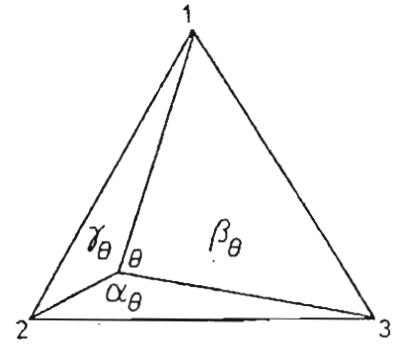

Fig. 16.

is given by

$$0(1) = \mu(\alpha_\theta),$$
$$0(2) = \mu(\beta_\theta),$$
$$0(3) = \mu(\gamma_\theta).$$

Now suppose our single observation x is equal to 1, and let us use this picture to describe the allocation of probability corresponding to PL_1. The key is to think of M as the collection of our probability masses, and to think of α_θ as the probability mass than can get into the singleton $\{\theta\}$. Hence

$$\bigvee_{\theta \in A} \alpha_\theta$$

is the total probability mass that can get into $A \in 2^\Theta$, and

$$PL_1(A) = \mu\left(\bigvee_{\theta \in A} \alpha_\theta\right).$$

The probability mass $\bigvee_{\theta \in A} \alpha_\theta$ is shown for several different finite subsets in Figure 17.

Now consider a finite subset A of Θ, and consider the probability mass

$$\bigwedge_{\theta \in A} \alpha_\theta.$$

This will be the total probability mass that can get into each and every point of A. Its measure will be the commonality number for A:

$$(33) \qquad Q_1(A) = \mu\left(\bigwedge_{\theta \in A} \alpha_\theta\right).$$

Can we compute $Q_1(A)$ from our geometric picture?

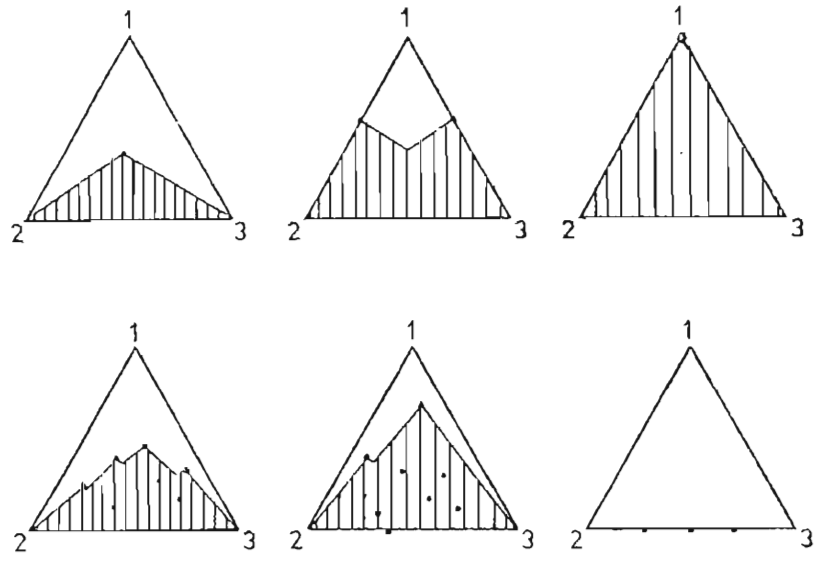

Fig. 17. In each example, the elements of A are represented by dots, and the probability mass $\bigvee_{\theta\in A}\alpha_\theta$ is shaded. In the last example each of the three elements θ of A has $\theta(1)=0$ and hence lies on the base of the triangle; the probability mass $\bigvee_{\theta\in A}\alpha_\theta$ is therefore empty.

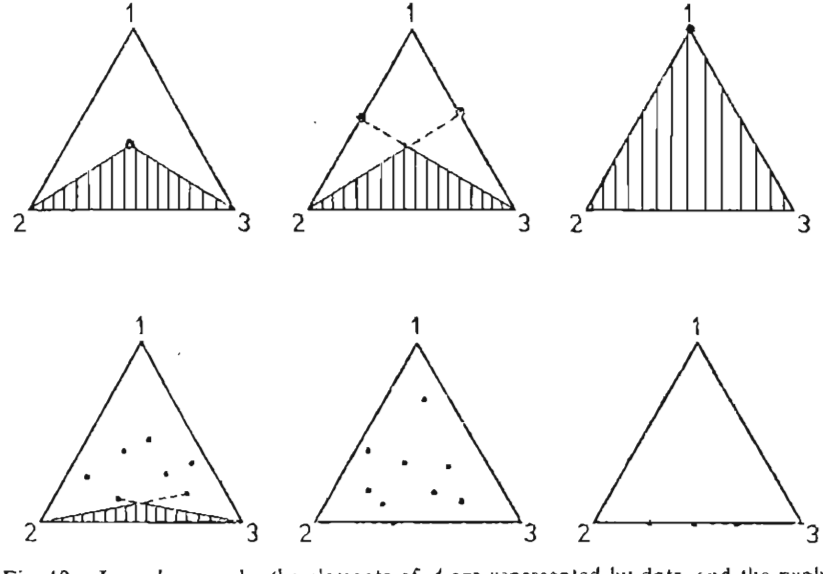

Fig. 18. In each example, the elements of A are represented by dots, and the probability mass $\bigwedge_{\theta\in A}\alpha_\theta$ is shaded. In the last two examples, there is an element θ of A with $\theta(1)=0$; hence the probability mass $\bigwedge_{\theta\in A}\alpha_\theta$ is empty.

The probability mass $\bigwedge_{\theta\in A}\alpha_\theta$ is shown for a few finite sets A in Figure 18. From these pictures, it is evident that we need to consider two cases: the case where $\theta(1)=0$ for some $\theta\in A$, and the case where $\theta(1)>0$ for all $\theta\in A$. In the first case,

$$Q_1(A)=\mu\left(\bigwedge_{\theta\in A}\alpha_\theta\right)=0. \tag{34}$$

In the second case, there will be a unique point θ_A of the triangle such that

$$\alpha_{\theta_A}=\bigwedge_{\theta\in A}\alpha_\theta, \tag{35}$$

$$\frac{\theta_A(1)}{\theta_A(2)}=\min_{\theta\in A}\frac{\theta(1)}{\theta(2)}, \tag{36}$$

and

$$\frac{\theta_A(1)}{\theta_A(3)}=\min_{\theta\in A}\frac{\theta(1)}{\theta(3)}. \tag{37}$$

(θ_A may be one of the points of A, as in the first and third examples of Figure 18, or it may not be, as in the second and fourth examples there.) From (33) and (34), we find that

$$Q_1(A)=\mu\left(\bigwedge_{\theta\in A}\alpha_\theta\right)=\mu(\alpha_{\theta_A})=\theta_A(1).$$

Can we compute $\theta_A(1)$ from (36) and (37)?

We can. Inverting these two relations gives

$$\max_{\theta\in A}\frac{\theta(2)}{\theta(1)}=\frac{\theta_A(2)}{\theta_A(1)}$$

and

$$\max_{\theta\in A}\frac{\theta(3)}{\theta(1)}=\frac{\theta_A(3)}{\theta_A(1)}.$$

Of course,

$$\max_{\theta\in A}\frac{\theta(1)}{\theta(1)}=\frac{\theta_A(1)}{\theta_A(1)}=1;$$

$$\sum_{y\in X}\max_{\theta\in A}\frac{\theta(y)}{\theta(1)}=\sum_{y\in X}\frac{\theta_A(y)}{\theta_A(1)}=\frac{1}{\theta_A(1)},$$

nd

$$(38)\qquad Q_1(A)=\theta_A(1)=\frac{1}{\sum_{y\in X}\max_{\theta\in A}\frac{\theta(y)}{\theta(1)}}.$$

Notice that (34) and (38) do indeed agree with (31).

Several of the properties of PL_1 should be evident from this geometric epresentation. Condensability, for example, means that the area of $\vee_{\theta\in A}\alpha_\theta$ can always be approximated arbitrarily closely by the area of $\vee_{\theta\in A'}\alpha_\theta$ for some finite subset subset A' of A. The second postulate of plausibility is satisfied, for whenever

$$\frac{\theta_1(1)}{\theta_1(2)}\geqslant\frac{\theta_2(1)}{\theta_2(2)}\quad\text{and}\quad\frac{\theta_1(1)}{\theta_1(3)}\geqslant\frac{\theta_2(1)}{\theta_2(3)},$$

we have $\alpha_{\theta_2}\leqslant\alpha_{\theta_1}$, whence

$$PL_1(\{\theta_1,\theta_2\})=\mu(\alpha_{\theta_1}\vee\alpha_{\theta_2})=\mu(\alpha_{\theta_1})=PL_1(\{\theta_1\}).$$

And the first postulate of plausibility is also satisfied, for

$$PL_1(\{\theta\})=\mu(\alpha_\theta)=\theta(1).$$

I have developed the geometric representation for $x=1$, but the analogous development for $x=2$ and $x=3$ should be obvious. We would have, for example,

$$PL_2(A)=\mu\left(\bigvee_{\theta\in A}\beta_\theta\right)$$

and

$$PL_3(A)=\mu\left(\bigvee_{\theta\in A}\gamma_\theta\right).$$

Finally, I should remark that an analogous geometric representation is possible when X has any number of elements. In this case, we have used a triangle, but in the general case we would use a $(k-1)$-dimensional simplex, where k is the number of elements in X. Hence the name 'simplicial'.

4.3.4. *The Continuous Case*

The simplicial method of defining plausibility functions can also be extended to the continuous case, though the extension is more problematic than in the linear case. The key to the extension is formula (32).

Suppose, indeed, that Θ_0 is a specification on a continuous space X. Then the division of X into discrete categories will result in commonality numbers

$$(39)\qquad Q_x^\gamma(A)=\frac{k_\gamma}{\sum_{g\in\gamma}\max_{\theta\in A}\frac{\int_g\theta(y)\,dy}{\int_{g_0}\theta(y)\,dy}},$$

where γ consists of all the categories and g_0 is the category containing x. As the categories are made finer and finer, we might expect (39) to tend to

$$(40)\qquad\frac{k}{\int\max_{\theta\in A}\frac{\theta(y)}{\theta(x)}\,dy}$$

for some constant k. Such a convergence does occur in many cases, and if the constant k is not zero, then the quantities defined by (40) can be taken as the commonality numbers for a plausibility function $PL_x\colon 2^{\Theta_0}\to[0,1]$.

5. The Historical Background

I have exposited the ideas in this essay as directly as possible, avoiding extensive historical references lest they evoke controversies or preconceptions. But I will now briefly review a few of the more prominent features in the historical background of these ideas.

5.1. Degrees of Belief

The notion of degree of belief has roots in the seventeenth and eighteenth centuries. The writers of that period usually used the terms 'degree of certainty' or 'degree of probability' rather than 'degree of support' or 'degree of belief', but they were concerned with the same questions and reached some of the same answers as I have discussed here. Notable

examples are James Bernoulli, in his *Ars Conjectandi* (published posthumously in 1713) and J. H. Lambert in his *Neues Organon* (1764). Both of these writers considered degrees of probability for the case of two alternatives that took the general form exhibited in Table II, but Lambert seems to have had the clearer view of the matter, for it was in the process of correcting Bernoulli's faulty and less general rule that he arrived at the rule described in §1.1.

The modern student of probability, whether statistician or philosopher, may find it strange to hear numbers like those in Table II called probabilities. For we have been indoctrinated in the view that probabilities must be additive. In other words, the numbers s_1 and s_2 in Table II are nowadays required to add to one before they can be called probabilities. But no such doctrine appears to have been known to Bernoulli or Lambert, for they discuss non-additive examples of probabilities without apology.

How then did the rule of additivity come to be applied to probabilities? The answer to this question is surely to be found in the fact that there are two kinds of probability. On the one hand there is the aleatory kind – the objective probabilities that are given by aleatory laws. And on the other hand there is the epistemic kind – the degrees of certainty, of support, of belief, or what-have-you. Bernoulli seems to have regarded the first kind as a special case of the second, and I took a similar view in §4.3. But in any case, the first kind incontestably must obey the rule of additivity; whereas, contrary to some contemporary opinion, it is highly doubtful that the second kind ought always to obey the rule of additivity. But it is easy to confuse and plausible to identify the two kinds of probability and thus apply the rules for aleatory probabilities to epistemic probabilities. Unfortunately, Laplace rather deliberately made precisely such an identification. Perhaps he wanted to deal with aleatory probabilities but found it suited his determinism to call them epistemic probabilities. Whatever his motivation, one effect of his great synthesis was the suppression, for a century and a half, of non-additive probabilities.

Indeed, it is remarkable how thoroughly Laplace's successors, and students of the history of probability down to this very day, have ignored Bernoulli's and Lambert's work on non-additive probabilities. Lambert's ideas were exposited by Prevost and Lhuilier in 1797, but since then they seem to have plunged into almost total obscurity. Issac Todhunter did discuss Prevost and Lhuilier's article in his *History* in 1865 (pp. 461–463), but apparently with little comprehension; and I know of no further reference to these ideas by any student or historian of probability during the past century.

5.2. Statistical inference

I have spoken in quite general terms in the preceding exposition, and I may have left the impression that my subject has been the whole field of statistical inference. But my subject has been much smaller than that. It has been the problem of statistical support – the problem of measuring the support for various subsets of Θ when a statistical specification $\{P_\theta\}_{\theta\in\Theta}$ is taken as known. In practice, the statistical specification may be far from known; and the problem of providing a specification – or of assessing the adequacy of a proposed one – seems to be the harder part of the theory of statistical inference.

Fifty years ago, R. A. Fisher distinguished between the problems of specification and the problems of 'estimation', declaring that the problems of specification were entirely 'a matter for the practical statistician'. For Fisher, it was the problem of estimating the correct value of θ on the basis of the specification $\{P_\theta\}_{\theta\in\Theta}$ and the observations $\mathbf{x}$ that was central to theoretical, as opposed to practical statistics. Today we would probably modify Fisher's judgment in several respects. First of all, his notion of estimation turned out to be too narrow, for the evidence about θ often cannot be summarized by an estimate and a measure of the estimate's accuracy. Hence many of us would replace the 'problem of estimation' with the 'problem of support' in a general description of the business of theoretical statistics. Secondly, we have come to see the problem of specification as a quite theoretical problem. As we have come to admit, the practical statistician's struggle with data is not only antecedent to the statistical specification – it is also antecedent to the specification of an 'observation space' X and even to the notion that anything is being governed by an aleatory law.

So in a modern view of theoretical statistics we might distinguish two broad problems – the problem of support and the problem of specification. The problem of support, which is the subject of this exposition is probably the more modest and manageable of the two.

5.3. Apologia

This essay is in part a defense and in part a reformulation of earlier work by A. P. Dempster. The ideas in it were inspired by my study of Dempster's work – a study that began when I attended his seminar on statistical inference at Harvard in the spring of 1971; and it culminates, in §4, with a justification of some of his methods of assessing statistical evidence. But, quite naturally, it differs in important respects from Dempster's work.

Most importantly, my philosophical account differs from Dempster's own. For far from rejecting the Laplacean synthesis, Dempster saw it as the source of 'much of the motivation and fascination of the modern science of probability'. (Research Report S-3, Harvard University, p. 8.) And instead of thinking of his lower probabilities as degrees of belief or degrees of support, he preferred, at least originally, to think of his upper and lower probabilities as bounds for some true but somehow unknowable probabilities, thus retaining the identification of degrees of belief with additive probabilities.

The mathematical account in §3 also differs from Dempster's earlier account. The differences derive mainly from the replacement of multivalued mappings by allocations of probability and from the isolation of the notion of condensability – innovations that permit the role of the commonality numbers to be fully developed.

Finally, the degrees of plausibility adduced in §4 overlap with but are not identical with the upper probabilities adduced in Dempster's papers. Those obtained by the simplicial method in §4.3 are identical with the upper probabilities produced by Dempster's 'structures of the second kind', (see p. 349 of his 1966 article.) Those obtained by the linear method in §4.2 are not discussed in any of Dempster's published work, though he has privately encouraged interest in the method. And the upper probabilities that result from Dempster's 'structures of the first kind' bear no relation to the present essay – the arguments given here provide no justification for them.

It is the new understanding of the meaning of Dempster's upper probabilities that I offer as the primary contribution of this essay. Since Dempster's own interpretation has not proven widely appealing, I hope that the more radical understanding of the present essay will inspire a wider interest.

Whether or not this hope is fulfilled, I must express my gratitude to my wife Terry and my many other friends, teachers and fellow students who have helped me with these ideas.

ACKNOWLEDGEMENT

This essay was written while I was supported by a graduate fellowship from the National Science Foundation, and it was revised while I was supported by contract N00014-67A0151-0017 from the Office of Naval Research.

Dept. of Statistics, Princeton University

REFERENCES

Bernoulli, James: 1713, *Ars Conjectandi*, Basel. Especially pp. 217–223. See also pp. 22–34 of the translation by Bing Sung, issued as Technical Report No. 2 of the Department of Statistics, Harvard University, February 12, 1966, and available in microfiche from the Clearinghouse for Federal Scientific and Technical Information, Washington, D.C.

Bernoulli, Daniel: 1777, 'The Most Probable Choice Between Several Discrepant Observations and the Formation Therefrom of the Most Likely Induction'. *Acta Acad. Petrop.*, pp. 3–33. Reprinted as pp. 157–167 of Pearson and Kendall's *Studies in the History of Statistics and Probability*, Griffin, 1970.

Choquet, Gustave: 1953–4, 'Theory of Capacities', *Annales de l'Institut Fourier, Université de Grenoble*, V, pp. 131–296.

Dempster, A. P.: 1966, 'New Methods for Reasoning Towards Posterior Distributions Based on Sample Data'. *Ann. Math. Statist.* 37, 355–374.

Dempster, A. P.: 1967, 'Upper and Lower Probabilities Induced by a Multivalued Mapping', *Ann. Math. Statist.* 38, 325–339.

Dempster, A. P.: 1968, 'A Generalization of Bayesian Inference (with discussion)', *J. Roy. Statist. Soc. Ser. B.* 30, 205–247.

Dempster, A. P.: 1968, *The Theory of Statistical Inference: A Critical Analysis*, Ch. 2, Research Report S-3, Dept. of Statistics, Harvard University. September 27, 1968.

Fisher, R. A.: 1922, 'On the Mathematical Foundations of Theoretical Statistics', *Philosophical Transactions of the Royal Society of London, Series A*, Vol. 222, pp. 309–368. Reprinted in R. A. Fisher, *Contributions to Mathematical Statistics*, Wiley, New York, 1950.

Lambert, Johann Heinrich: 1760, *Photometria*, Augsburg, pp. 131–147.

Lambert, Johann Heinrich: 1764, *Neues Organon*, Zweiter Band, pp. 318–421. Reprinted in 1965 as Vol. II of *Lambert's Philosophische Schriften* by Georg Olms Verlagsbuchhandlung, Hildesheim.

Lehman, E. L.: 1959, *Testing Statistical Hypotheses*, Wiley, New York.

Prevost and Lhuilier: 1797, 'Mémoire sur l'application du Calcul des proba-

bilités à la valeur du témoignage', *Mémoires de l'Académie Royale de Berlin*, pp. 120-152.

Shafer, Glenn: 1973, *Allocations of Probability: A Theory of Partial Belief*. Doctoral dissertation submitted to the Dept. of Statistics, Princeton University, June 26, 1973. Available from University Microfilms, Ann Arbor, Michigan.

Todhunter, Issac: 1865, *A History of the Mathematical Theory of Probability*. Reprinted by the Chelsea Publishing Co., New York, 1949.

DISCUSSION

Commentator Good: Suppose the evidence arises in two experiments with a thousand heads in one and a thousand tails in the other. Would the order in which the data are obtained effect the results?

Shafer: No, the order does not matter. Dempster's rule of combination is symmetric.

Lindley: This theory owes much to Dempster's work. My own view of that theory is that it is upset by Aitchison's counter-example. (This was presented in the discussion to Dempster's 1968 paper: *J. Roy. Statist. Soc. B*, 30, 205–247 on page 234.) Essentially Aitchison considers two trinomial distributions with probabilities (0.4, 0.5, 0.1) and (0.4, 0.1, 0.5) – the essential point being the equality of the probabilities for the first class. He then shows that, according to Dempster's theory, a single observation falling in the first class can change our opinions about which trinomial obtains. This is very counter-intuitive since the observation would appear to contribute nothing to this question. My question to tonight's speaker is: does the same criticism apply to his theory?

Shafer: Aitchison's criticism applies to the simplicial method, which is indeed the same as the method in Dempster's 1968 paper. The criticism does not apply to the linear method.

Perhaps I should take a paragraph to cast Aitchison's example in the vocabulary of the preceding essay. The statistical specification consists of two aleatory laws: $\Theta = \{\theta_1, \theta_2\}$. The set of possible outcomes is, say, $X = \{\text{Azure, Brown, Crimson}\}$. Both aleatory laws assign Azure a chance of 0.4, but they disagree on Brown and Crimson. One begins, presumably, with the vacuous belief function: $\text{Bel}(\{\theta_1\}) = \text{Bel}(\{\theta_2\}) = 0$. But what support function over Θ should one have after a single observation $x = \text{Azure}$? (1) Following the simplicial method, one obtains the support function S over Θ given by $S(\{\theta_1\}) = S(\{\theta_2\}) = \frac{2}{9}$. (2) But following the linear method, one obtains $S(\{\theta_1\}) = S(\{\theta_2\}) = 0$ – i.e., there is no change from the vacuous support function with which one began.

Choosing between the linear and simplicial methods in this example

obviously amounts to deciding whether the observation x = Azure should be treated as no evidence (linear solution) or as precisely balanced conflicting evidence (simplicial solution). In defense of the simplicial solution, one might argue that the single observation x = Azure is indeed internally conflicting: the observation of Azure rather than Crimson supports θ_1 while the observation of Azure rather than Brown supports θ_2. But this is not very convincing, and I now agree with Mr Aitchison and Mr Lindley that the linear solution is preferable.

Writing in January of 1975, I should point out that the original essay and talk that inspired Mr Lindley's question argued for the simplicial method and did not mention the linear method. The essay printed above, which gives equal billing to the two methods, was written in the summer of 1973. My present preference for the linear method is based primarily on its adaptation to the notion of 'weight of evidence', a notion which allows a much deeper understanding of Dempster's rule of combination. (See my forthcoming book *A Mathematical Theory of Evidence*.)