# Statistical Evidence and Belief Functions

Teddy Seidenfeld

University of Pittsburgh

In his recent monograph [7], Professor Shafer has offered us an alternative to Bayesian inference with his novel theory of belief functions and, in his current paper [8], has characterized his position by pointing to two basic differences it shares with Bayesianism. First, belief functions are non-additive so that the degree of belief assigned to the disjunction '$A_1$ or $A_2$' may be larger than the sum of the degrees of belief assigned to the separate disjuncts. Second, the theory of belief functions has its own rule for determining the commitments to changes in degrees of belief when evidence is compounded. So, instead of the Bayesian postulate of conditionalization, that is, in place of using Bayes' theorem to identify the commitments to changes in probability when evidence accumulates, the theory advocated by Professor Shafer relies on a proposal he traces to A.P.Dempster, which he calls Dempster's rule for combination of belief functions. In my comments here I want to focus attention on the second of these proposals, the replacement of conditionalization by the combination rule, and I hope to argue that there is a serious defect in the theory of belief functions because of this replacement.

Before engaging in that criticism, however, I would like to point out that much of Professor Shafer's attack against the Bayesian position, an attack that serves as motivation for his own program, is relevant to the first of the two basic points at issue and is not relevant to the second. That is, when I object to the rule of combination (and suggest it is inferior to conditionalization) I am not thereby constrained from echoing many of the very same worries about Bayesianism that Professor Shafer raises in his contribution to this symposium.

PSA 1978, Volume 2, pp. 478-489


https://doi.org/10.1086/psaprocbienmeetp.1978.2.192484 Published online by Cambridge University Press

Let me illustrate this. Suppose we are faced with a strangely bent coin, about which we are relatively ignorant, and are asked to determine our degree of belief in the proposition that, when next flipped, the coin will land heads-up. As proper Bayesians we are obligated to specify a precise probability function that assigns some value $\underline{p}$, $0 \leq \underline{p} \leq 1$, to this proposition and thereby we assign a value $1-\underline{p}$ to the contradictory proposition that the coin will not land heads-up. Professor Shafer reminds us of the old, but relevant, problem that if the magnitude of $\underline{p}$ signifies the strength of our belief that the coin will land heads-up, then we cannot use zero (for $\underline{p}$) to represent our ignorance, since that assignment leads to a maximally strong belief that the coin won't land heads-up. But we profess ignorance about the coin's tendency to land one way as opposed to some other.

An all too familiar Bayesian response to this challenge consists in conceding that point and solving the problem by adopting a version of the Laplacean principle of Insufficient Reason. Admitting that the magnitude of the probability cannot be interpreted to reflect the weight of the evidence supporting the proposition (that is, agreeing that we cannot argue ignorance is equated with no evidence is equated with no support), some Bayesians try to find a way out in terms of the <u>symmetries</u> of the precise probability function used to represent ignorance.[1] Thus, just in case $\underline{p} = 1/2$, the probabilities are equal for the two possibilities: that the bent coin lands heads-up and that it doesn't.

Equally familiar are the two objections that rebut this answer. (1) The solution is aprioristic since it conflates the case of ignorance with that of significant background knowledge. For instance, if the only serious possibilities for an outcome of the flip are landing heads-up or landing tails-up, then the equal probability assignment fails to distinguish ignorance from the assumption that the coin is <u>fair</u>. (2) The solutions generated by symmetry considerations are inconsistent with the probability calculus. For instance, if the serious possibilities for an outcome of the flip include landing heads-up, landing tails-up, and landing on edge, then a blind application of the equal probability rule (to capture ignorance) results in a probability of 1/2 for each of the three alternatives when, separately, each is compared to its contradictory, e.g., landing heads-up or not landing heads-up. This second objection becomes very serious when the problem is of a conventional statistical sort with a continuum of basic alternatives. A uniform (equal probability) distribution with respect to one parameterization of the continuum is not uniform with respect to an equivalent non-linearly transformed parameterization.

I bother to rehearse this old problem with you only because the inadequacies of fiducial inference are of a kind with this Bayesian quandary over how to represent ignorance in a precise probability function. The fiducial argument was R.A.Fisher's recipe for solving the inverse inference problem: inference from observed "sample" to unobserved "population", without the use of Bayesian ingredients, such as a Bayesian "prior" probability for representing ignorance. Leonard Savage characterized Fisher's ploy as an attempt to make the Bayesian omelette without breaking the Bayesian eggs: an attempt to have a precise posterior probability without admitting a precise prior probability. We can easily understand what goes wrong with Fisher's fiducial argument by noting that the serious paradoxes surrounding this mode of inference (paradoxes involving simple one parameter problems) stem from alternative, mutually incompatible representations of ignorance.

It is for reasons like these that I welcome the first break with Bayesian theory found in Professor Shafer's belief functions. By using non-additive measures we can assign a probability of zero to each alternative: landing heads-up, landing tails-up, etc.; yet we note that some one outcome must eventuate by assigning the value 1 to the disjunction of alternatives. I welcome this break with Bayesian theory for, relying on a reinterpretation of Professor Shafer's theory in terms of Dempster's original position, the non-additive property of belief functions translates into the non-additive property of <u>lower</u> probability measures when the agent's beliefs are represented by <u>intervals</u> of probability. What is gained over the Bayesian position is the use of <u>sets</u> of probability functions to represent a belief state. Thus, ignorance is properly represented by the [0,1] interval, whose lower bound is 0.

Unfortunately, Professor Shafer does not subscribe to this reading of his position, and I am at a loss to fully understand why. However, the challenge I want to raise against Shafer's program in no way depends upon this interpretation of the non-additive feature of his system; for it is the adequacy of the combination rule (the substitute for Bayesian conditionalization) which is the subject of my comments.

This symposium is titled <u>Statistical Evidence</u> and I think it most appropriate that we consider the application of the theory of belief functions to problems of statistical inference. In his contribution to this session [5], Professor Levi has demonstrated for us the danger in trying to tame Fisher's fiducial argument according to Dempster's strategy. As I understand his analysis, we capture the beast at the expense of losing simple direct inference; that

https://doi.org/10.1086/psaprocbienmeetp.1978.2.192484 Published online by Cambridge University Press

is, we must forfeit Professor Ian Hacking's *Frequency* principle.[2] No doubt, then, Professor Shafer shows wisdom when he withdraws from Dempster's safari in search of the elusive Fisher. Also, we see from Levi's presentation that even to accommodate direct inference about a "pivotal" variable, Shafer's program would need alterations to handle unusual frames of discernment, i.e., unusual partitions that include the data-to-be-acquired in the frame. Thus, the target of my criticism is the more mundane variety of statistical problem discussed by Shafer in his book ([7], chapter 11), simple inverse inference.

Following his account, let us grant ourselves the liberty of chances, or aleatory probability (as Shafer calls it). That is, in addition to the epistemic measure of support, belief functions, there is also aleatory probability. In chapter 11 of *A Mathematical Theory of Evidence*, Shafer fixes the connection between the two concepts in a "convention for assessing statistical evidence" ([7], p. 238).

Let me paraphrase his rule. Suppose we are faced with a simple inverse statistical problem, we have flipped the bent coin once and noted the outcome. The statistical hypotheses, binomial hypotheses that the coin is biased with an aleatory probability $\theta$, $0 \leq \theta \leq 1$, form the *frame of discernment*, i.e., the ultimate partition of interest here. The statistical evidence is the report of the outcome of the trial. The *support* for a (composite) hypothesis $\underline{A}$, that a subset A of these basic possibilities includes the true bias, is given by the formula:

$$S_x(\underline{A}) = 1 - [\underline{\max}_{\theta \varepsilon \bar{A}}\, p_\theta(x) \div \max_{\theta}\, p_\theta(x)] \tag{1}$$

where: 'x' stands for the statistical evidence; '$\bar{A}$' stands for the complement of A (in the parameter space); and '$p_\theta(x)$' stands for the aleatory probability (chance) of x if $\theta$ is the true statistical bias. In more familiar terms, the support for the disjunction $\underline{A}$ of simple statistical hypotheses (each disjunct is called a "singleton" by Shafer) is a function of the maximum likelihood *of* the complement of A: one minus the maximum likelihood of $\bar{A}$. Hence, the support for $\underline{A}$, given data x, cannot be high unless $\underline{A}$ includes all those singletons of high (relative) likelihood.

When the data are compound, as when the evidence consists of several observed trials, there are two avenues open for determining support. The repeated trials may be treated as a single compound trial and the rule for support (1) may be applied once (where the entire evidence determines the likelihood function). Alternatively, elementary support functions may be calculated from each datum separately and an overall support function determined by applying Dempster's combination rule to put the simple support functions together.

Dempster's rule is the rule in Shafer's theory for combining distinct belief functions. Put very roughly, if $B_1$ and $B_2$ are two belief functions based on distinct bodies of evidence and if the two families of hypotheses involved in these functions are suitably related (Shafer describes this as a common frame of discernment), then the orthogonal sum of $B_1$ and $B_2$ is the new (combined) belief function: written as $B_1 \oplus B_2$.[3]

As Professor Shafer notes, the alternative routes for determining a support function (using compound statistical evidence) do not result in the same belief function ([7], chapter 11, §3). Let me borrow Professor Shafer's own illustration of this point. Say, for simplicity, we know that the bent coin is either biased .9 for landing heads (and .1 for landing tails)--call this hypothesis $\theta_1$--or it is biased .3 for landing heads (and .7 for landing tails)--call this hypothesis $\theta_2$. Suppose we flip the coin twice and observe a head on the first toss--$x_1$--and a tail on the second--$x_2$.

If the data are treated as an outcome of a compound trial, flip twice, then the convention for determining support yields the numbers:

$$\begin{aligned} S_{(x_1,x_2)}\theta_1 &= 0 \\ S_{(x_1,x_2)}\theta_2 &= 4/7. \end{aligned} \tag{2}$$

If the data are decomposed into the two elementary events, $x_1$ and $x_2$, the convention (1) used twice to generate two simple support functions, $S_{x_1}$ and $S_{x_2}$, and then Dempster's rule used to combine these into a single, compound support function, $S_{x_1} \oplus S_{x_2}$, the resulting numbers are:

$$\begin{aligned} [S_{x_1} \oplus S_{x_2}]\theta_1 &= 2/9 \\ [S_{x_1} \oplus S_{x_2}]\theta_2 &= 6/9. \end{aligned} \tag{3}$$

Shafer reacts to this situation with two remarks. First, he suggests that the solution which uses the combination rule applied to the simple support functions has the advantage of being sensitive to the "conflict among the observations" ([7], p. 250), which the solution based on the compound trial suppresses. (This "conflict" is a species of Shafer's notion of "dissonance".) Second, he points out

https://doi.org/10.1086/psaprocbienmeetp.1978.2.192484 Published online by Cambridge University Press

the existence of another measure, called plausibility, defined as:

$$Pl_x(\underline{A}) = \max_{\theta\varepsilon A} p_\theta(x) \div \max_{\theta} p_\theta(x) = 1 - S_x(\bar{\underline{A}}), \qquad (4)$$

preserves relative plausibility for singletons no matter which method of solution is adopted ([7], p. 250). For instance, in the foregoing example the relative plausibility of $\theta_1$ to $\theta_2$ is 3:7 for both methods.

I will respond to each of these remarks with the assistance of several examples. Let me begin with a statistical problem that has achieved some notoriety in current literature.[4] Suppose we are faced with an inverse statistical inference involving one parameter, the correlation $\rho$, in a bivariate normal distribution with known means and variances. For simplicity, we may think of the problem as one concerning the correlation between the errors $(x_i, y_i)$, where $x_i$ is the error in the i-th reading with unbiased instrument X (whose errors are distributed normally with unit variance) and similarly $y_i$ is the error in the i-th reading with unbiased instrument Y (whose errors are also distributed normally with unit variance). Put succinctly, $(x_i, y_i)$ is a pair from the bivariate normal distribution with $\mu_x = \mu_y = 0$, $\sigma_x^2 = \sigma_y^2 = 1$, and unknown correlation $\rho$. Moreover, separate pairs are statistically independent.

Suppose the compound data are n pairs: $(x_1, y_1), \ldots, (x_n, y_n)$. We plead ignorance about $\rho$, at first, so our initial support function is the vacuous one that assigns 0 to any non-tautologous hypothesis. Consider, next, the partition of the data into two distinct parts:

$$x = (x_1, \ldots, x_n) \text{ and } y = (y_1, \ldots, y_n).$$

Since x has a distribution that is free of the unknown correlation, we can quickly identify the support function $S_x(\rho)$. In fact, it is the vacuous suppport function. That is, learning x tells us nothing about $\rho$. With perfect symmetry, y has a probability distribution free of the unknown correlation, so the support function $S_y(\rho)$ is, again, vacuous with respect to the parameter of concern, $\rho$. That is, learning y tells us nothing about the correlation. The two support functions, $S_x(\rho)$ and $S_y(\rho)$, are based on separate data (that collectively exhaust all the data). Applying Dempster's combination rule yields $[S_x \oplus S_y]\rho$, which is the vacuous support function (see footnote 3). We begin in ignorance and, after using this method to evaluate the new evidence available, we remain in ignorance. [Note that further subdivision of the data, say into 2n components: $x_1, y_1, \ldots, x_n, y_n$, leads to the same result.]

https://doi.org/10.1086/psaprocbienmeetp.1978.2.192484 Published online by Cambridge University Press

Alternatively, we may treat the entire data as a single compound trial, i.e., n observations from the bivariate normal distribution, and construct a support function $S_{(x,y)}\rho$ by using the convention, formula (1). This procedure duplicates the technique demonstrated in the example of the bent coin for using a compound trial, which led to the support function $S_{(x_1,x_2)}\theta$. However, the support function and plausibility function generated this way for inference about the correlation are non-vacuous. By taking the compound data as a unit and applying (1), we produce the non-vacuous support function $S_{(x,y)}\rho$ whose plausibility function is fixed by the likelihood function:

$$(1-\rho)^{-n/2}\cdot\exp[-(U-2V\rho)/(2[1-\rho^2])], \qquad (4)$$

where $U = \Sigma_i(x_i^2 + y_i^2)$ and $V = \Sigma_i(x_i y_i)$. That is, from (4) we learn a lot about $\rho$. But the data are the same as in the preceding analysis.

How can we explain these conflicting accounts? The data are, by one procedure, irrelevant yet, by the other, they are relevant to the inverse inference about $\rho$. The aspect of the combination rule that accounts for this phenomenon is that Dempster's rule relies on <u>products</u> of simple support functions and, from a Bayesian point of view, that makes sense as long as the bits of evidence used to determine the simple support functions (to be combined) are <u>statistically</u> independent. The combination rule ignores all the conditional distributions of one datum given another. More formally, the combination rule builds joint distributions from marginal distributions and that is a dangerous technique.

With the partition of the bivariate data into the x and y parts, all the relevant information about $\rho$ is contained in the conditional distribution of x given y, or y given x. There is no relevant information about $\rho$ contained in x or in y taken separately. Individually, each is an <u>ancillary</u> statistic, one whose probability distribution is free of the parameter of interest for inverse inference. It is an elementary consequence of conditionalization (or even of the Likelihood Principle) that ancillary data are, by themselves irrelevant. Thus, from a Bayesian (or Likelihoodist) perspective, there is no fault to be found in convention (1) as it applies in this example. It <u>is</u> correct to say that x (or y) alone tells us nothing about $\rho$. What is deficient is the combination rule used with arbitrary partitions of composite data, even when the partitions generate perfectly accurate simple belief functions (as in this illustration).

It is a mistake to think that the only risk we run in

https://doi.org/10.1086/psaprocbienmeetp.1978.2.192484 Published online by Cambridge University Press

applying the combination rule to "unusual" partitions is a potential loss of information. The example just discussed has the feature that all the pertinent information is lost when the individual likelihoods are combined by Dempster's rule. However, we can fool ourselves into thinking there is more in a given body of data just as easily by relying on the combination rule, and this is my reason for rejecting Professor Shafer's account of "conflict" (or "dissonance"). Suppose we flip the bent coin n times. Instead of considering just partitions into statistically independent components, we may build up very long lists of complicated summaries, such as: the percent of heads showing in subsequences of every j-th flip; the percent of heads showing in the first m trials, $1 \leq m \leq n$, etc. Generally, each such report will generate a different support function (by convention (1)) and we can use Dempster's rule to combine the lot of them. Since it might appear that each new summary captures a new, relevant feature of the compound data (with respect to inverse inference about the coin's bias), applying the combination rule to these many support functions should expose ever new aspects of conflict within the data. If there is an advantage to be found in methods that expose conflict, then the more intricate the family of support functions we can build from a given body of evidence, the merrier the analysis we obtain by combining them according to Dempster's rule.

One of Fisher's basic teachings is the lesson that there are limited amounts of new information that can be extracted from a given body of statistical evidence (with respect to a given problem of inverse inference). Assuming either conditionalization, or even a simple likelihood principle, we see that in some cases the entire data may be summarized in a single sufficient statistic.[5] Though two aspects of the same data may lead to different support functions when considered separately, one may be irrelevant given the other if the latter is sufficient. For example, with inverse inference about a binomial distribution (the statistical model for the process of flipping a bent coin), the pair of numbers, number of flips, percent of flips landing heads-up, are jointly sufficient for inference about the binomial parameter, in place of the whole data. (Note that, if the number of flips is ancillary, the percent of flips landing heads-up is conditionally sufficient, i.e., exhaustive, for the data.) All other features of the composite sample, features reflected in the "conflict" among the parts of the data, are irrelevant given the sufficient statistics, if conditionalization is valid.

In his book Professor Shafer qualifies his discussion of statistical inference by limiting it to problems where the data are partitioned into "physically independent observa-

https://doi.org/10.1086/psaprocbienmeetp.1978.2.192484 Published online by Cambridge University Press

tions" ([7], p. 238). We see now that he means by that restriction a division into statistically (or aleatorily) independent observations. In common statistical parlance, the rule of combination works with i.i.d. trials, identically independently distributed trials, and it fails with statistically dependent trials. In his book Professor Shafer qualifies the application of the combination rule to belief functions based on "distinct" bodies of evidence. The qualification is repeated in sections 3 and 8 of [8]. I take it that "distinct" means non-redundant in the way that all other aspects of the evidence are redundant once a sufficient statistic is known.[6]

In section 8 of [8], Professor Shafer questions the applicability of Bayes' theorem, i.e., the practicality of conditionalization, for fear that we might not be able to find a formulation of our problem that permits us to determine the requisite probabilities for calculating according to Bayes' rule. At one point he suggests that "it will not be plausible to regard the $E_i$ [the partition of the evidence] as conditionally independent given A and $\bar{A}$" ([8], p. 458), and he uses this suggestion to question one's ability to obey conditionalization. But if it is a practical difficulty for Bayesians to work with conditionally dependent observations, and I do not see those cases requiring new computational skills, it is a theoretical prohibition that prevents Professor Shafer from using them in his own program. At least the Bayesian theory provides the machinery for deciding whether the data are mutually independent. How does the theory of belief functions resolve the general problem of distinctness of the evidential bits? How am I to know whether, in the example of section 4 in [8], the evidence $E_3$ (that the stubs in the attic suggest a roof over the concrete section) is independent of, i.e., non-redundant with, $E_4$ (that my neighbor's testimony as to the original use of the house suggests a concrete floor)? If it can be argued that dependence obtains between these two, how then do I reformulate the evidence so that the combination rule applies?

I have tried to show that the novel theory of belief functions is to be applauded where it departs from Bayesian theory by using non-additive measures interpreted as lower bounds on sets of probabilities, but it is to be reproached where it departs from Bayesian theory by substituting the combination rule for conditionalization. I have tried to argue that the combination rule is unsatisfactory because of its limitation to partitions of the data that are statistically independent if the problem is aleatory, and because of its limitation to partitions of the data that, in general, are free of redundancy (in the sense of sufficiency). The principle of conditionalization is not subject to either of these restrictions. Moreover, it provides the Bayesian with criteria for judging the adequacy of particular applications of Dempster's rule.[7]

Lest we forget, two contemporary philosophers have been hard at work, one for nearly two decades, developing alternatives to Bayesianism that, like Professor Shafer's program, are free of the requirement that only precise probability functions serve to represent a rational agent's belief state. Henry Kyburg, as does Shafer, rejects conditionalization in addition. His theory, <u>epistemological probability</u> (see [3]), comes closest among the competitors I am aware of to a reconstruction of Fisherian statistics. My concerns with his position are over the extent to which conditionalization fails epistemologically; specifically, sufficiency is invalid epistemologically. (So Fisher is on the loose once again!)

More recently, Isaac Levi has constructed a theory of <u>indeterminate probability</u> (see [4]) that preserves conditionalization within convex sets of coherent probabilities. However, not much of the Fisherian project (nor contemporary statistics, for that matter) survives intact in Levi's theory. That does not show his program is wrong, for it is not obvious to me that classical statistics is worth saving. It does suggest that any inductive logic sophisticated enough to treat ignorance respectably may be too mature to accept the naive statistical view of what it means not to know anything of relevance. Above all, we should agree that in inverse inference ignorance is anything but bliss.

## Notes

[1] I take it that this criticism is what Shafer reports at the bottom of page 460 in [8].

[2] This conclusion follows from Levi's argument on page 469 of [5].

[3] Two interesting properties of Dempster's rule (demonstrated by Shafer in chapter 3 of [7]) are worth mentioning here. First, if either belief function, say $B_1$, is the vacuous one (that represents ignorance) then its combination with any other belief function $B_2$ leaves $B_2$ unaltered, i.e., we see that $B_1 \oplus B_2 = B_2$. Second, the combination rule is invariant over the order in which the belief states are combined, i.e., $B_i \oplus B_j = B_j \oplus B_i$.

[4] Barnard and Sprott attribute it to Basu [2] in their [1].

https://doi.org/10.1086/psaprocbienmeetp.1978.2.192484 Published online by Cambridge University Press

In fact, it can be traced to Savage ([6], p. 20).

[5]A statistic s is sufficient for data d, with respect to a parameter of interest $\theta$ (for inverse inference) just in case: $p(d/\theta \& s) = p(d/s)$.

[6]In inverse (statistical) inference a statistical model is assumed. Where the question arises whether the data are to count as evidence (as in testing for "outliers") and where the question arises what statistical model is to be accepted (as in sample "re-use" methods), convention (1) does not apply since we lack a well defined "frame of discernment".

[7]Thus, I reject Shafer's claim ([8], pp. 459 and 464) that conditioning is a special case of Dempster's rule. Also, if his remarks (page 459 of [8]) are intended as a response to my questions about the adequacy of the combination rule (left unaided by conditionalization for fixing the conditions of "distinctness"), then those remarks appear to me to be beside the point.

## References

[1] Barnard, G. and Sprott, D.A. "A Note on Basu's Examples of Anomalous Ancillary Statistics." In Foundations of Statistical Inference. Edited by V.P. Godambe. Montreal: Holt, Rinehart and Winston, 1970. Pages 163-170.

[2] Basu, D. "Recovery of Ancillary Information." Sankhya A. 26 (1964): 3-16.

[3] Kyburg, H. The Logical Foundations of Statistical Inference. Dordrecht: Reidel, 1974.

[4] Levi, I. "On Indeterminate Probabilities." The Journal of Philosophy LXXI(1974): 391-418.

[5] -------. "Dissonance and Consistency According to Shackle and Shafer." In PSA 1978. Volume 2. Edited by P.D. Asquith and I. Hacking. East Lansing, Michigan: Philosophy of Science Association, 1981. Pages 466-477.

[6] Savage, L. "Subjective Probability and Statistical Practice." In The Foundations of Statistical Inference. Edited by L. Savage et al. London: Methuen, 1962. Pages 9-35.

[7] Shafer, G. A Mathematical Theory of Evidence. Princeton: Princeton University Press, 1976.

[8] ---------. "Two Theories of Probability." In PSA 1978. Volume 2. Edited by P.D. Asquith and I. Hacking. East Lansing, Michigan. Philosophy of Science Association, 1981. Pages 441-465.