Journal of Scientific Exploration, Vol. 8, No. 4, pp. 491–508, 1994
0892-3310194

Applied Scientific Inference

P. A. STURROCK

Center for Space Science and Astrophysics, ERL 306, Stanford University, Stanford, CA 94305-4055

Abstract–The purpose of this article is to use the principles of scientific inference to provide guidance in evaluating complicated issues such as those raised by the study of anomalous phenomena. Specifically, the article presents a formalism (a "protocol") for organizing and combining the many judgments that must be made in the scientific evaluation of the relevant hypotheses. All judgments are to be expressed as probabilities, and the rules for combining probabilities are derived from Bayes' theorem. Setting up a problem in a manner that permits such an analysis can be helpful in imposing a structure and discipline upon the analysis, and also in exposing relevant questions that might otherwise have remained hidden. Furthermore, the introduction of probabilities makes it possible to put on a sound numerical basis such assertions as "extraordinary claims require extraordinary evidence." One finds that extraordinary evidence can be built up from many (but not very many) items of unspectacular evidence, provided the items are truly independent.

The proposed procedure permits a clear separation between a statement of initial prejudice and an evaluation of the significance of considered evidence. However, it would be even better to set out explicitly the considerations on which the prejudice is based, and to view those considerations as part of the evidence to be evaluated. The procedure also draws a clear separation between the roles and judgments of data analysts (who assign probabilities to specified statements, based on the evidence), and those of theorists (who assign probabilities to the same statements, based in turn on the considered hypotheses).

In order to reach a consensus on any topic, it is recommended that probability estimates be made by teams of experts, all team-members being presented with the same data but acting independently, and procedures are proposed by which individual estimates may be combined to yield a consensus estimate.

1. Introduction

"Formality has its place in guiding one along suitable paths of argument: and most of us need some guidance." (D. V. Lindley: see Lindley, Tversky and Brown 1979, p. 177)

"Talking of a Court-martial that was sitting upon a very momentous publick occasion, he expressed much doubt of an enlightened decision; and said, that perhaps there was not a member of it, who in the whole course of his life, had ever spent an hour by himself in balancing probabilities." (Samuel Johnson 1780)

In most areas of mainstream science, the "research protocol," a term that we use to denote the basic rules and procedures of the research process, is normally so well defined and so generally understood that it does not merit discussion, much less debate. However in the study of anomalous phenomena, such as those discussed in the pages of this journal, the research protocol may not be well defined and so may not be generally understood, and may in consequence be wide open for debate.

There is a temptation to dismiss, sometimes with prejudicial terms, topics for which the research protocol is not well defined. For instance, by labeling a topic as "paranormal," one is not too subtly implying that only "normal" phenomena occur, so that the topic in question cannot really happen and therefore must be bogus. The premises for that decision are usually not spelled out. Such an attitude leads to a severe limitation in the scope of scientific research, a limitation that might be acceptable to some scientists but is not acceptable to all scientists. Most Members of SSE have some interest in some topics that do not now belong to mainstream science. For most of these topics, determining the most appropriate research protocol is one of the principal problems to be addressed.

The main goal of this article is to propose a specific protocol for the study of topics that now lie outside of mainstream science. The protocol appears to be appropriate and workable, but perceptive scientists will no doubt identify defects in this proposal, and this step should lead to improved protocols that are more reliable and more powerful.

In searching for such a protocol, one is led to look for the basis on which science is constructed. One school of thought takes this basis to be the discipline of scientific inference. The procedures of scientific inference, as expounded for instance by Jeffreys (1931) and Good (1950), are typically based on Bayes' theorem that is the algorithm showing how ones' assessment of the probability of an hypothesis should be updated by reference to one's evaluation of the relevant data. Another very clear exposition of the role of probability theory in science and engineering is contained in a series of lectures given by Jaynes (1959). Bayes' theorem will be presented very briefly in Section 2. In its usual form, Bayes' theorem is not too convenient for the evaluation of hypotheses in scientific research, although it does provide the basis of a very powerful procedure for data analysis known as the Maximum-Entropy Method. (See, for instance, Gull 1988, Skilling 1959.) Section **3** briefly presents an evaluation scheme (Sturrock, 1973), based on Bayes' theorem, that has been found to be more convenient for application to real scientific problems.

An important concept in the theory of scientific inference is that of "prior probability," that represents one's assessment of an hypothesis before the relevant evidence is examined. There is more than one way to take account of one's initial assessment or "prejudice." This topic is discussed in Section 4, and it is argued that communication among investigators is facilitated by regarding prior evidence simply as one more block of evidence, on a par with all other relevant evidence that is to be produced by and shared among investigators.

A significant shortcoming of the usual procedures of scientific inference is that all decisions are implicitly made by one person. Such procedures cannot be reconciled with precepts concerning the scientific process that have been advocated, for instance, by Ziman (1978). According to Ziman, the scientific process is essentially the development of a rational consensus concerning the problem under investigation. One way to reconcile these two positions, the scientific-inference position and the rational-consensus position, is to modify the former in such a way that all assessments are made by consensus rather than by an individual. Section 5 presents a procedure by which this might be accomplished. It is not claimed that this is the best procedure, merely that it is a plausible and workable procedure. Since the combination of many distribution functions can be difficult if carried out analytically, or time-consuming if carried out computationally, we discuss in Section 6 how reasonable estimates of the same quantities might be made by Monte-Carlo methods. Further thoughts on these questions and procedures are offered in Section 7.

2. Bayes' Theorem

We use the notation $P(A|B)$ to denote the probability that the proposition A is true on the basis of knowledge that proposition B is true. As usual, we adopt the convention that the measure of probability extends over the range zero to unity: $P(A|B) = 0$ if A is impossible given B, and $P(A|B) = 1$ if A is certain given B.

The notation AB stands for the logical product of the two propositions A and B. Then AB is true if and only if both A and B are true. The product rule of probability theory (Good 1950) states that

$$P(AB|C) = P(A|BC)P(B|C). \tag{2.1}$$

Since AB = BA,

$$P(A|BC) = P(B|AC)P(A|C). \tag{2.2}$$

Hence

$$P(A|BC) = \frac{(B|AC)}{P(B|C)} P(A|C). \tag{2.3}$$

This is usually referred to as Bayes' Theorem.

A change in notation clarifies the significance of this theorem. In writing

$$P(H|NZ) = \frac{P(N|HZ)}{P(N|Z)} P(H|Z), \tag{2.4}$$

the symbols have the following interpretation: H is an hypothesis under consideration, Z represents the "baseline" or "zero-base" information, and N represents a new item of information. (Note that, in a real situation, Z will normally incorporate assumptions that are so familiar that they are unrecognized. Yet some of these assumptions may prove, in the course of analysis, to be incorrect.) Then $P(H|Z)$ is the "prior probability" and $P(H|NZ)$ is the "post probability;" $P(N|Z)$ is the probability that N is true, based only on the baseline knowledge Z; and $P(N|HZ)$ is the probability that N is true, evaluated on the assumption that the baseline knowledge Z and the hypothesis H are both true. The quantity $P(N|\mathrm{HZ})$ is referred to as the "likelihood" of N, referred to the hypothesis H, for given evidence Z.

We see that, if N is likely to be true on the basis of Z alone, knowledge that N is true will not greatly increase the probability of H. On the other hand, the probability of H will be increased significantly if N is unlikely on the basis of Z alone, but likely on the basis of both Z and H.

If N is impossible (or highly unlikely) on the basis of Z and H, the fact that N is true makes H impossible (or very much less likely).

Note that if we set $P(H|Z) = 0$, then $P(H|\mathrm{NZ}) = 0$, no matter what the new evidence may be. Similarly, $P(H|NZ) = 1$ if $P(H|Z) = 1$, for any N. Hence one must avoid assigning probabilities zero or unity to any proposition (unless it is logically impossible on the basis of the given information), since this entails that we can never change these values, no matter what subsequent information may turn up. For further comments on this point, see, for instance, Good (1950, p. 49).

If we introduce the notation $\bar{\mathrm{H}}$ for "not H," we see from (2.4) that

$$\frac{P(H|NZ)}{P(\bar{H}|NZ)} = \frac{P(N|HZ)}{P(N|\bar{H}Z)} \frac{P(H|Z)}{P(\bar{H}|Z)}. \tag{2.5}$$

The ratio on the left-hand side is called the "odds" on H, based on the information N and Z. Hence (2.5) shows that the post-odds is equal to the prior-odds multiplied by a quantity that is the ratio of the likelihoods of N based on H and on H.

An even more useful concept is that of "log-odds," defined by

$$\Lambda(H|Z) = \log_{10} \frac{P(H|Z)}{P(\bar{H}|Z)}. \tag{2.6}$$

We then see that (2.5) may be written as

$$\Lambda(H|NZ) = \Lambda(H|Z) + W(H{:}N|Z), \tag{2.7}$$

where, following Good (1983), we have introduced the notation

$$W(H{:}N|Z) = \log_{10}\left(\frac{P(N|HZ)}{P(N|\bar{H}Z)}\right) \tag{2.8}$$

for the logarithm of the ratio of the likelihoods. Equation (2.7) shows that the post-log-odds is obtained by adding to the prior-log-odds the logarithm of the likelihood ratio. Good (1983) refers to the latter as the "weight" of the evidence for H provided by the evidence N, given the evidence Z. The term inside the logarithm is referred to, by Good (1983), as the "Bayes-Jeffreys-Turing factor."

It is convenient to use the symbol A for the log-odds, in order to reserve the symbol L for the log-likelihood, to be introduced in Section 5.

Equation (2.7) demonstrates the usefulness of the scientific-inference approach to controversial topics. Even if two investigators begin with very different prior beliefs, they should be able to agree upon the significance of a new block of evidence N. Hence they should be able to agree on whether the new evidence makes the hypothesis more or less likely and (as measured by log-odds), they should be able to agree upon the amount by which the hypothesis becomes more or less likely.

One can readily see from (2.7) that if the "new" information is comprised of several blocks of independent information, $N_{\gamma,} \gamma = 1,2,...,G$, the final log-odds is given by

$$\Lambda(H|N_1N_2...N_GZ) = \Lambda(H|Z) + W(H:N_1|Z) + W(H:N_2|Z) + ... + W(H:N_G|Z). \quad (2.9)$$

The concept of "decibel" is familiar and useful in electrical engineering. It is also a useful term to use in scientific inference. The measure of log-odds in db is 10 times the log-odds. Hence an increase or decrease in odds by a factor of 2 corresponds to a change in log-odds by 3 db or–3db, respectively; an increase in odds by a factor of 100 corresponds to 20 db; etc. We may therefore express our prejudice concerning a proposition in terms of db, and we may express our assessment of evidence also in terms of db.

These concepts enable us to revisit the familiar assertion that "extraordinary claims require extraordinary evidence" (Sagan, 1994). An extraordinary claim would be an hypothesis with a low prior probability corresponding, say, to a log-odds of-6, or-60 db. In order to convert this to an odds of 100:1, corresponding to a log-odds of 2, or 20 db, we clearly need 80 db of evidence. If this evidence were to be derived from a single case, it would indeed be extraordinary. However, we see from (2.9) that the same weight of evidence could be derived from eight cases, each with log-odds of only 10 db, or from sixteen cases, each with log-odds of only 5 db. Hence a combination of many (but not very many) cases, each of which is unspectacular, can yield evidence equivalent to one quite extraordinary case. Of course, it is essential that the cases be completely independent, and that the source of each case be sufficiently credible. One should clearly require that different cases come from independent sources since, if they all came from the same source, one would be forced to require that that source be credible at the 80 db level, that in many instances would be asking rather a lot.

We here note only briefly that, in case analysis, the weight of evidence of given information, as measured by the log-odds, will require at least two judgements: the significance of the case material, taken at face value; and the credibility of the source or sources. This topic was touched on briefly in Sturrock (1993), and will be developed further in a later article.

3. The Interface Scheme

In science, the evaluation of an hypothesis involves an interplay between empirical evidence on the one hand, and theoretical work on the other hand. The evidence may be obtained from laboratory experiments, from observations, or from interviews of witnesses and other field investigations. We assume that a data analyst (who may or may not be the experimenter, observer, or field investigator) has responsibility for inspecting and summarizing the evidence, and a theorist has responsibility for analyzing the hypotheses, in such a way as to facilitate the evaluation of theoretical hypotheses on the basis of the evidence.

We formalize the relationship in terms of a model in which there is a set of statements, relevant to the problem in hand, that are comprehensible to both data analysts and theorists. These statements form the "interface" between the data-analysis and theoretical activities. We further require that all statements of the interface are to be arranged in groups, each group comprising an "item." For present purposes, we assume that the items are logically independent of each other. That is to say, the assessments of one item do not depend logically upon the assessments of another item. We assume that there is a finite set of items **I„** α = 1,2, . . . ,**A,** and that each item **I,** comprises a finite set of statements s„, k = 1,2, . . . ,K_α, so chosen that the statements are mutually exclusive and form a complete set. That is, for an item such as I_α, it is logically demonstrable that one and only one of the statements $S_{\alpha k}$ is true. Properties of this model will here be discussed only briefly. For a more complete description, see Sturrock (1973).

In this model, a data analyst (or team of data analysts) and a theorist (or team of theorists) communicate only by separately assigning probabilities to the statements of each item. The data analyst will assign a probability $P(S_{\alpha k}|EZ)$ to each statement $S_{\alpha k}$, based on the relevant evidential data E and the baseline data Z. The theorist must consider not a single hypothesis, but a complete set of mutually exclusive hypotheses H,, i = 1,2, . . . ,**I,**so that it is logically demonstrable that one and only one of the hypotheses H_i is true. The task of the theorist is to assign a probability $P(S_{\alpha k}\ H_iZ)$ to each statement S„, based on each hypothesis H_i and the baseline data Z. In dealing with a real scientific problem, it may not be possible to specify all relevant hypotheses explicitly. If H,,H_2, . . . ,H_I is a set of hypotheses which have been identified and are mutually exclusive, and for each of which a theory may be developed, but which do not form a complete set, we may complete the set by adding the "something else" or "ignorance" hypothesis H_0. We are to admit complete ignorance concerning the consequences of H_0, and we are to consider that H_0, H,, H„ . . . ,H_I comprise a complete and mutually exclusive set of hypotheses. With this understanding, the probabilities $P(S_{\alpha k}|H_0Z)$, for each item I„ are to be chosen to be maximally noncommittal, subject to the restrictions imposed by the baseline data Z. Procedures for assigning probabilities that are maximally noncommittal, subject to prescribed constraints, have been developed for application to the maximum-entropy method of data analysis. (See, for instance, Jaynes 1968, Gull 1988, Skilling 1989.)

The present model is shown schematically in Figure 1. It can be shown (Sturrock, 1973) that if communication between the data analyst and the theorist occurs only through the interface of this model, and if they consider only a

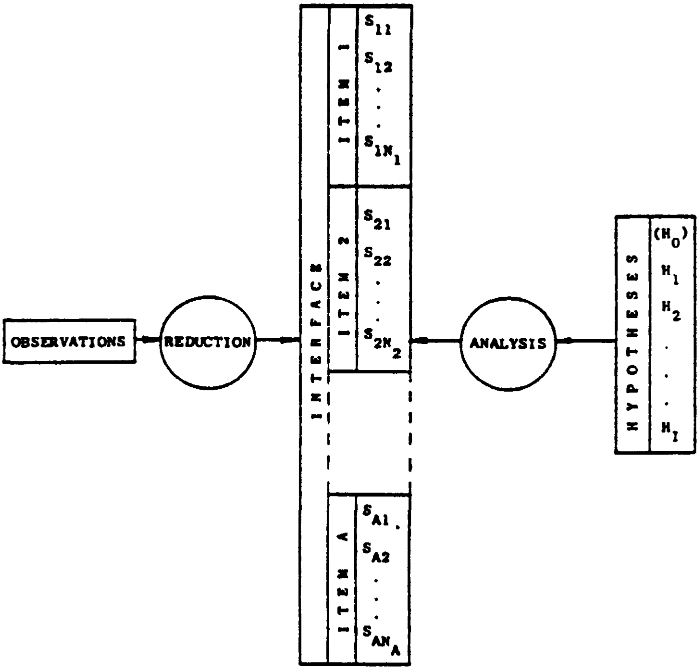

Fig. 1. Schematic representation of model used for evaluation of relationship between theory and observations.

single item, then the post probabilities of the hypotheses are related to the prior probabilities by the following formula,

$$P(H_i \mid F_\alpha(E)Z) = \left[\sum_k \frac{P(S_{\alpha k} \mid H_i Z)P(S_{\alpha k} \mid EZ)}{\sum_j P(S_{\alpha k} \mid H_j Z)P(H_j \mid Z)} \right] P(H_i \mid Z), \qquad (3.1)$$

where we introduce the notation $P(H_i \; F_\alpha(E)\, Z)$ to denote the post-probability of H_l based on only one item (or "fact") F_α of evidence E.

If the A items enumerated by $\alpha = 1,2, \ldots ,A$, are taken to be independent, then the probabilities of the hypotheses, based on all items, may be obtained by combining the probabilities obtained (by means of equation (3.1) for each item according to the following formula:

$$P(H_i \mid EZ) = \frac{P(H_i \mid F_l(E)Z) \ldots P(H_i \mid F_A(E)Z)\left[P(H_i \mid Z)\right]^{-(A-l)}}{\sum_j P(H_j \mid F_l(E)Z) \ldots P(H_j \mid F_A(E)Z)\left[P(H_j \mid Z)\right]^{-(A-l)}}. \qquad (3.2)$$

It is clear that the sum of these probabilities is unity, and that the final estimate is independent of the order in which the facts are combined.

In scientific research concerning a specified topic, one attempts to assemble independent blocks of evidence. Suppose that the total evidence E can be broken down into several independent blocks that we denote by E„ γ = 1,2, . . . ,G. Then the probabilities of the hypotheses, based on all the blocks of evidence, is given by the following formula that is analogous to (3.2):

$$P(H_i \mid E_1 ... E_G Z) = \frac{P(H_i \mid E_1 Z) ... P(H_i \mid E_G Z)\left[P(H_i \mid Z)\right]^{-(G-1)}}{\sum_j P(H_j \mid E_1 Z) ... P(H_j \mid E_G Z)\left[P(H_j \mid Z)\right]^{-(G-1)}}. \qquad (3.3)$$

4. Priors and Prejudice

In probability theory, it is conventional to refer to the probability assessment made in advance of assessment of the evidence as the "prior probability," and to assume that this estimate fully represents the "prejudice" of the analyst. If we were to follow this course, and if there were only a single analyst, the prejudice of the analyst would be represented by the values of $P(H_i|Z)$.

This approach raises problems for the approach now being proposed. For instance, we ask the data analyst and the theorist to assign probabilities $P(S_k|$EZ) and $P(S_k|H_iZ)$ to each statement S_k, based on their assessment of the evidence E and on their analysis of each hypothesis H_i. This implicitly assumes that the data analyst and the theorist have the same baseline data (the same "prejudice"), but this is unlikely to be the case. Furthermore, we wish to proceed to consider that judgments are made not by a single data analyst and a single theorist, but by a team of data analysts and a team of theorists.

The principal problem is usually that different investigators have very different prior beliefs concerning the hypotheses under investigation. In discussing UFO evidence, for instance, one investigator may consider it quite plausible that extraterrestrial beings exist and are visiting Earth, whereas another investigator might consider that proposition to be ludicrous. Indeed, it is this large spectrum of prior beliefs that is responsible for much of the difficulty in discussing such issues.

We may therefore simplify the evaluation procedure by isolating the prejudice of each analyst concerning the hypotheses under consideration. We introduce the symbol E_0 to represent the prior information that any investigator may have that has a recognized influence upon his prior assessment of the probabilities of the hypotheses. (Here and elsewhere, "his" is used as an abbreviation for "his or her," etc.) We now denote by Z the remaining (implicit) information that, by definition, has no recognized influence upon the prior probabilities. Then equation (3.3) will be replaced by

$$P(H_i \mid E_0 E_1 ... E_G Z) = \frac{P(H_i \mid E_0 Z) P(H_i \mid E_1 Z) ... P(H_i \mid E_G Z)\left[P(H_i \mid Z)\right]^{-G}}{\sum_j P(H_j \mid E_0 Z) P(H_j \mid E_1 Z) ... P(H_j \mid E_G Z)\left[P(H_j \mid Z)\right]^{-G}}. \qquad (4.1)$$

Since information *Z* has no recognized influence upon the prior probabilities, it would in many cases be reasonable to assign equal probabilities to the members of any discrete complete set of hypotheses, as determined by the information *Z* alone. The equation (4.1) would then simplify to the special case

$$P(H_i \mid E_0E_1...E_GZ) = \frac{P(H_i \mid E_0Z)P(H_i \mid E_1Z)...P(H_i \mid E_GZ)}{\sum_j P(H_j \mid E_0Z)P(H_j \mid E_1Z)...P(H_j \mid E_GZ)}. \tag{4.2}$$

This approach has certain advantages:

A. The data analyst is not required to state his prejudice concerning the hypotheses, but he is required to ignore his prejudice in carrying out his work. Different data analysts should be able to agree on the probabilities $P(S_n|EZ)$, even though they may have very different prejudices.

B. Each theorist is required to begin his work by representing his prejudice concerning the hypotheses by the set of probabilities $P(H_i|E_0Z)$. This would hopefully lead him to do some soul-searching to determine exactly what E_0 comprises. He may find that, if E_0 is set out explicitly, he will give rather different values to the probabilities $P(H_i|E_0Z)$ than he would have done if E_0 had remained unstated and un-analyzed.

C. Once the theorist has stated the values he assigns to $P(H_i|E_0Z)$, all further assessments of probabilities must be made independently of his prejudice E,. Different theorists should be able to agree on the probabilities $P(S_n|H_iZ)$, even though they may have widely divergent estimates of the probabilities $P(H_i|E_0Z)$.

D. One has the option of simply presenting the significance of the **evidence** concerning the hypotheses without ever inquiring into the prejudice of the analysts. If this is done, it is a simple matter for any interested person to combine the probabilities $P(H_i|E_1 \ . \ . \ . \ E_GZ)$, based purely on the evidence, with probabilities $P(H_i|E_0Z)$ that represent his personal prejudice:

$$P(H_i \mid E_0E_1...E_GZ) = \frac{P(H_i \mid E_0Z)P(H_i \mid E_1...E_GZ)\left[P(H_i \mid Z)\right]^{-1}}{\sum_j P(H_j \mid E_0Z)P(H_j \mid E_1...E_GZ)\left[P(H_j \mid Z)\right]^{-1}}. \tag{4.3}$$

5. The Joint Evaluation of Probabilities

We now wish to consider the case that the evidence is being analyzed by a team of data analysts (or by several teams, one for each item, or one for each block of evidence), and the theoretical assessments are being made by a team of theorists. For each assessment that must be made, there will be not a single evaluation of the probability $P(A|B)$ of statement A on the basis of statement B, but a number of such evaluations. We need to find a procedure by which a "manager," acting on behalf of the team, may choose a probability that is a fair representation of the collection of opinions that have been expressed in terms of individual probabilities. This interesting question has previously been addressed by Good (1979) and by Lindley, Tversky and Brown (1979).

It clearly would be unsatisfactory to try to adopt a single number that is supposed to represent the collection of assessments. If all members of a team agree exactly in their evaluations of the probability, that value should be given much greater weight than, say, the average of a set of widely divergent probabilities that represent a wildly varying set of assessments. As a minimum increase in complexity, we should require that the collective evaluations be represented by two numbers, just as one will normally represent a set of measurements of a physical variable by the mean and the standard deviation of those measurements.

We will then need to have some set of rules for combining the summary representation of two sets of evaluations. These rules should lead to the same result as the summary representation of all the original assessments taken together. Since the summary representation of a set of measurements in terms of a mean and a standard deviation satisfies this requirement, it clearly is tempting to try to apply this representation of measurements to the summary representation of a set of probabilities.

However, it would not be appropriate to treat the actual probability assessments as the basic data that are analogous to a set of measurements of a physical variable, such as the position of a mark on a line, or the time of occurrence of an event. For these physical measurements, we could begin with the assumption that all values of the coordinate measurement, from $-\infty$ to $+\infty$, are equally likely. This clearly is not the case for a probability, that is necessarily restricted to the range 0 to 1. Following Lindley, Tversky and Brown (1979), it is here proposed that the appropriate datum to work with is the log-odds Λ , introduced in Section 2 and defined by (2.6). We recall from Section 2 that we may combine two assessments, based on two independent blocks of evidence, simply by adding the appropriate log-odds.

Consider the situation that there are two investigators who have very different prejudices represented by their prior probabilities for a proposition, who nevertheless agree upon the significance of an empirical datum, as represented by its probability assessment. Each investigator will change his log-odds by precisely the same amount. This is similar to the situation of two scientists interested in the value of a physical quantity who begin with very different estimates of that quantity, but who agree that analysis of new information implies that earlier assessments require correction by being increased (or decreased) by an agreed amount. Hence there are some similarities here with the situation concerning the measurement of a physical variable such as position or time.

Now suppose that members of a team of N investigators, $n = 1,2, \ldots ,N$, consider a proposition **A,** in the light of evidence B, and make *independent* assessments $\Lambda_n(A|B)$ of the log-odds of A, given B. In order to provide a convenient representation of these assessments, we agree to compare them with a simple distribution of log-odds, and adopt the Gaussian form:

$$D(\Lambda)d\Lambda = \frac{d\Lambda}{(2\pi)^{1/2}\sigma} e^{-(\Lambda-\mu)^2/2\sigma^2}. \tag{5.1}$$

As a rationale for adopting the Gaussian distribution, we offer the following argument. Suppose that a large number of analysts are required to evaluate statement *A* on the basis of statement B. However, suppose that these analysts are also subjected to many small items of information. Each item of information will lead to a small change $\Delta\Lambda$ in the log-odds Λ that the analyst is assigning to the proposition. Note that $\Delta\Lambda$ is independent of A. It is shown in the Appendix that these assumptions lead to a Gaussian form for the log-odds distribution.

We now need to select a guideline for evaluating the parameters of the "parent" Gaussian distribution from the estimates made by the team members. A reasonable prescription is to adopt the maximum-likelihood procedure (Brandt, 1976) for determining the two parameters–the mean μ and the standard deviation σ. The log-likelihood, defined by

$$L \equiv \ln\left[\prod_{n=1}^{N} D(\Lambda_n)\right], \tag{5.2}$$

yields

$$L = \sum_{n=1}^{N} \ln\left[\frac{1}{(2\pi)^{½}\sigma} e^{-(\Lambda_n-\mu)^2/2\sigma^2}\right]. \tag{5.3}$$

One readily finds that the values of μ and σ that maximize L are given by

$$\mu = \frac{1}{N}\sum_{n=1}^{n} \Lambda_n, \tag{5.4}$$

and

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} (\Lambda_i - \mu)^2. \tag{5.5}$$

It may be verified that one will obtain the same estimates of the mean and standard deviation by considering all data together, or by first dividing the data into two sets, representing each set by a mean and standard deviation, and then using (5.4) and (5.5) to combine estimates for the two data sets.

A useful extension of this approach is to give each investigator the option of representing his assessment of a log-odds by a distribution, rather than by a single number, since some investigators may be more confident of their assessments than are other investigators. Thus investigator n may represent his assessment by a distribution $D_n(\Lambda_n)d\Lambda_n$ that could conveniently be taken to be expressible in the form (5.1), with μ and σ replaced by μ_n and σ_n. By making an appropriate modification of the definition (5.3) of the log-likelihood function, we find that (5.4) and (5.5) are now modified to read

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \mu_n, \tag{5.6}$$

and

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}\left[\sigma_n^2 + \left(\mu_n - \mu\right)^2\right]. \tag{5.7}$$

In the scheme developed in the previous section, we are drawing a distinction between the evaluation of the prior probabilities, that we assume is based on the personal prejudices (that we refer to as E,) of the analysts, and the evaluation of all other probabilities, that we assume is based on the analysis of documented evidence (E,,etc.) or on the theoretical analysis of hypotheses. This can lead to a difficult situation. For instance, suppose that nine analysts proposed values of log-odds that were compatible with zero, with a standard deviation of unity, but a tenth analyst proposes a value of 1,000. A commonsense approach is to ignore an estimate that is so far out of line with the others. One could formalize this by deciding, for instance, to view with suspicion any estimate of a log-odds that in magnitude exceeds some value, for instance 10. The manager could, for example, decide to weight all estimates by the factor

$$W(\lambda) = \exp[-(\Lambda/10)^2]. \tag{5.8}$$

An alternative procedure is to adopt the limited goal of developing a consensus only from the opinions that are reasonably consensible. This limited goal would justify the manager in ignoring extreme expressions of opinion. Since it is dangerous to make such decisions during the evaluation process, he should state his policy at the outset. The following is proposed as a possible procedure.

The given set of estimates of the log-odds A are to be ordered. Of this ordered set, the manager will discard the top fraction f (10%, for instance) and the bottom fraction f, and attempt to provide a representation only of the central fraction 1-2f . We assume that this central section runs from the lower value L, to the upper value L,. We now need to devise a procedure for representing this central data set by a mean and standard deviation. This procedure must of course take account of the fact that the data set has been truncated.

A possible procedure is to form the cumulative distribution function C_e(L) such that C_e(L) is the number of estimates (from the truncated set) for which $L_n <$ L. We then form the normalized function C_{Ne}(L), defined for all values of L, by

$$\begin{aligned} C_{Ne}(\Lambda) &= 0, & \Lambda &< \Lambda_l, \\ C_{Ne}(\Lambda) &= N_e^{-1}\, C_e(\Lambda), & \Lambda_l &< A < A_{„}, \\ C_{Ne}(\Lambda) &= 1, & \Lambda &< \Lambda_u. \end{aligned} \tag{5.9}$$

where N_e is the number of estimates in the truncated set. We then compare this with the normalized cumulative distribution function for a truncated Gaussian distribution function that we define by

$$\begin{aligned} &C_g(\theta) = 0, && \theta < -\phi, \\ &C_g(\theta) = \frac{2erf(\theta) - f}{1-2f}, && -\phi < \theta < \phi, \\ &C_g(\theta) = 1, && \theta < \phi, \end{aligned} \tag{5.10}$$

where ϕ is chosen so that

$$2erf(-\phi) = f, 2erf(\phi) = 1-f. \tag{5.11}$$

Then we may estimate μ and σ by minimizing the integral

$$I = \int_{-\infty}^{\infty} d\theta \left[C_{Ne}(\mu + \sigma\theta) - C_g(e)\right]^2 \tag{5.12}$$

or, equivalently,

$$I = \int_{-\infty}^{\infty} d\lambda \, \sigma \left[C_{Ne}(\Lambda) - C_g\left(\frac{\Lambda - \mu}{\sigma}\right)\right]^2. \tag{5.13}$$

As suggested in point B at the end of Section 4, there are real advantages in putting the role of prior information on the same footing as that of all other information. Within the context of the present section, this could be achieved by giving a designated panel (the "Prior Panel") responsibility for compiling *evidence* that should determine the values of the priors. We could then regard E_0 as representing the explicit output of that panel, rather than the implicit prejudice of the theoretical analysts. It is then up to a panel of data analysts to convert the evidence E_0 into estimates of the log-odds of the various hypotheses, just as they do for E,,etc. It would be interesting indeed to see just what document would emerge from a panel charged with collecting and presenting evidence that could be used to evaluate priors for the various hypotheses that need to be considered in assessing evidence related to parapsychology, UFO reports, etc.

6. Sampling and Monte-Carlo Processes

In attempting to apply the procedures outlined above to a real problem, we would face the task of determining the distribution of the final probabilities $P(H_i|E_0E_1 \ . \ . \ . \ E_AZ)$, on the basis of the distribution of prior probabilities $P(H_i \ E_0Z)$, the distribution of probabilities $P(S_{\alpha k} \ H_iZ)$ of statements based on each hypothesis, and the distribution of probabilities $P(S_{\alpha k} \ E_\gamma Z)$ of statements based on each block of evidence. It may be possible to derive a comprehensive formula for the output distribution function, but that does not seem likely. Furthermore, even if it were possible, the formula is likely to be unmanageably complicated. An alternative approach is to perform the calculations for appropriately weighted ***sample*** values of the basic quantities $P(H_i|E_0Z)$, $P(S_{\alpha k}|H_iZ)$ and $P(S_{\alpha k}|F_k(E_\gamma)\,Z)$. For instance, one could run the calculation for all the pairs of equally weighted values $\mu - \sigma$ and $\mu + \sigma$, or one could run the calculation for all the triplets of equally weighted values $\mu - (3/2)^{1/2}\sigma$, μ and $\mu + (3/2)^{1/2}\sigma$.

If the problem involves a very large number of basic quantities, it will not be feasible to apply the above procedure to the entire calculation. One possibility is to use a Monte Carlo approach, running the calculation for a smaller number of cases that are chosen randomly, for instance selecting randomly from the pairs $\mu - \sigma$ and $\mu + \sigma$, or from the triplets $\mu - (3/2)^{1/2}\sigma$, μ and μ + (312)"o. Another approach is to proceed one step at a time. For each item, one could Sample two, three (or more) values of each probability, and estimate the corresponding probabilities of the hypotheses based on that item. One could then estimate, by the maximum-likelihood method of Section 5, log-likelihood distributions for the hypotheses based on that item. One can proceed to use the same procedure to estimate the log-likelihood distributions of all the items that comprise a block of evidence, and then repeat the procedure to estimate the log-likelihood distributions of all blocks of evidence combined.

In this context it should be noted that sampling the log-likelihood distributions for a complete set of hypotheses will not necessarily yield a set of probabilities that sum to unity. It seems reasonable to interpret the estimates arrived at in this way as "weights" of the hypotheses, and to derive estimates of the probabilities by normalizing the weights to sum to unity.

7. Discussion

The goal of this article is to facilitate the analysis of ill-defined scientific problems for which the hypotheses are vague and the data are less than secure. Although the exploration of a topic may be an individual enterprise, the procedures by which a consensus is achieved necessarily comprise a collective process. Furthermore, it is necessary to divide the overall process into separate activities that can be undertaken by experts in those activities. In astrophysical research, for instance, there is a clear separation between the roles of observers and theorists. In fact these roles are often further subdivided so that a more complete listing would include observers, data analysts, those who compile catalogs, model builders (who suggest specific hypotheses), those who analyze these models analytically, and those who analyze these models by means of computer simulations. It is likely that the study of anomalous phenomena would advance much more rapidly if research in these areas could be similarly subdivided. In astrophysics, one looks with great suspicion on an article by a single author who claims to have made new observations and also to have devised a theory to explain those observations. One should look with similar suspicion on an article that claims to do as much for an anomalous phenomenon.

In the hard sciences, there is the advantage that one is, for the most part, dealing with numerical data measurements made by the instruments appropriate to that area of research. In dealing with field investigations of such transient events as those that lead to UFO reports, one should clearly seek every opportunity to make detailed physical measurements, such as may be possible in investigating ground traces, for instance. Nevertheless, most investigations comprise interviews with witnesses, and these do not lend themselves to mensuration in the ordinary sense of the word. Hence the discipline introduced into a subject by the use of applied scientific inference can be useful in generating probability estimates for specific statements. Once information is reduced to numerical form, it becomes grist for the mill of numerical analysis that is subject to the power of modem computers.

A further advantage of applied scientific inference is that one can avoid loaded terms that generate more heat than light, and instead work with a non-prejudicial terminology that is acceptable to all investigators involved in the research enterprise. Terms such as "paranormal" and "pseudoscience" are, for our purposes, useless and must be replaced with more specific statements such as "the present laws of physics are complete," or "no event can occur that is not compatible with the present laws of physics," or "all properties of living systems can in principle be explained in terms of the properties of non-living systems." And rather than use terms such as "believer" and "skeptic," one will instead record a degree of belief in any statement by a probability – or, even better, by a distribution of log-odds. (After all, the only difference between a "believer" and a "skeptic" is what he believes, and how strongly.) This has the same advantage as replacing the statement "I am a strong tennis player" with the statement "I have won a city tournament, but not a county tournament," or replacing the statement "I am hard up" with the statement "I am $2,000 in the red."

Even if one does not enter all the probabilities in a worksheet, it is still instructive to draw it up. One must specify the precise hypotheses under consideration, one must list the relevant items of evidence, and flesh out the items by a set of precise statements. By going through this process, one will find that a thorough analysis of a problem requires posing and answering many questions that would otherwise have remained hidden.

In applying scientific inference to the study of anomalies, one will be led to devote more attention to theoretical questions. If one is simply being pressed to state what he has "proved," an investigator will be conservative, and this leads him to ignore the theoretical issues at stake. If, on the other hand, the study is subdivided, the theorist may accept his task of specifying in detail the relevant hypotheses, since he is not being asked to state what he has proved. What, if anything, is "proved" is the result of a collective process. After the hypotheses have been specified, the team of theorists will need to assess the prior probabilities. This activity also will be highly instructive, since it may be found that the priors for the explicit hypotheses are determined by the priors for more fundamental hypotheses that had remained concealed and unconsidered. (See, for example, Sturrock, 1993.)

The procedure outlined in this article can be extended in several ways. For instance, it is possible to deal with a continuous distribution of statements rather than a finite distribution (see, for instance, Sturrock, 1973). This extension is appropriate in dealing with measurements of continuous variables. It is also possible to introduce more than one interface in the analysis process. One interface may separate the raw data from the statements that are entered into a catalog. Another interface may separate the data compiled in the catalog from a few statements that summarize patterns that emerge from the catalog.

A necessary development will be the formalization of the relationship between the analysis of individual cases and the evaluation of "global" hypotheses that are relevant to the entire phenomenon. Hypotheses for individual cases will be of the type "this event was due to . . . ," whereas the global hypotheses will be of the type "some events are due to . . . " This development was touched on briefly in a recent paper (Sturrock, 1993).

Note that the procedures outlined in this article represent only one way to develop a theory of applied scientific inference that can be applied either to topics of mainstream science or to topics, such as anomalous phenomena, that are presently outside of the purview of mainstream science. It is to be hoped that this article will draw attention to the need for some such protocol, and that other improved procedures will be developed.

It should also be noted that a scientist has the option of becoming engaged or not becoming engaged in an investigation based upon the principles of scientific inference. A scientist can elect not to be a player. However, if a scientist makes this choice, but nevertheless wishes to express a judgment concerning an anomalous phenomenon, he then faces the challenge of proposing an alternative protocol for the study of such topics.

Concepts advanced in this article have previously been applied in brief communications concerning remote viewing (Sturrock, 1987) and the UFO problem (1993). These potential applications will be described in more detail in forthcoming articles.

Acknowledgments

I acknowledge with pleasure helpful comments by Slava Glukhov, I. J. Good and Carl Sagan.

Appendix

Asympototic Form for the Log-Odds Distribution Function

We now show that the following assumptions leads to a Gaussian asymptotic form for the log-odds distribution function. The investigator is subject to a uniform, steady stream of many small items of information, that we consider to be random.

It is convenient to introduce, for the purposes of this appendix, a time variable t. If we denote by ΔΛ the change in A due to one item of information, the distribution function $D(\Lambda,t)$ will satisfy the Fokker-Planck equation (Sturrock, 1994)

$$\frac{\partial D}{\partial t} = \frac{\partial}{\partial \Lambda}\left(\left\langle \frac{\Delta\Lambda}{\Delta t} \right\rangle D\right) + \frac{1}{2}\frac{\partial^2}{\partial \Lambda^2}\left(\left\langle \frac{(\Delta\Delta)^2}{\Delta t} \right\rangle D\right). \qquad \text{(A1)}$$

Since we are assuming that the inflow of random information is uniform, the coefficients in (A1) depend neither on Λ nor on t. Hence we may re-write (A1) as

$$\frac{\partial D}{\partial t} = -A\frac{\partial D}{\partial \Lambda} + \frac{1}{2}B\frac{\partial^2 D}{\partial \Lambda^2}, \qquad \text{(A2)}$$

where

$$A = \left\langle \frac{\Delta\Lambda}{\Delta t} \right\rangle, \; B = \left\langle \frac{(\Delta\Lambda)^2}{\Delta t} \right\rangle. \qquad \text{(A3)}$$

One finds that if the distribution has the form of a delta-function at A = 0 at time t = 0, the form at later times is given by

$$D(\Lambda,t) = \frac{1}{(4\pi Bt)^{\frac{1}{2}}}\exp\left[-\frac{(\Lambda - \mathrm{A}t)^2}{4Bt}\right], \qquad \text{(A4)}$$

that is seen to be Gaussian in form. Since the width of the function (A4) increases with time, it follows that any initially compact distribution will evolve asymptotically to the Gaussian form.

References

Brandt, S. (1976). *Statistical and Computational Methods in Data Analysis.* 2nd edition; Amsterdam: North-Holland.

Good, I. J. (1950). *Probability and the Weighing of Evidence.* London: Griffin.

Jaynes, E. T. (1959). Probability theory in science and engineering. Dallas, Texas: Socony-Mobil Oil Company Colloquium *Lectures in Pure and Applied Science,* No. 4.

Jaynes, E. T. (1968). *I.E.E.E. Transaction SCC,* 4,227.

Jeffreys, J. (1931). *Scientific Inference.* Cambridge University Press.

Good, I. J. (1979). J. *Statist. Computations and Simulations,* 9, 77.

Good, *I.* J. (1983). *Good Thinking: The Foundations of Probability and its Applications.* Minneapolis: University of Minnesota Press, 187.

Gull, S. F. (1988). *Maximum-Entropy and Bayesian Methods in Science and Engineering.* Eds. G. J. Erickson and C. R. Smith; Dordrecht, Kluwer, 1, 53.

Lindley, D. V., Tversky, A., and Brown, R. V. (1979). *J. Roy. Statist. Soc.,* Ser. A, 142, 146.

Johnson, Samuel (1780). In *The Life of Samuel Johnson L. L. D.* By James Boswell. New York: Random House, 915.

Skilling, J., Ed. (1989). *Maximum Entropy and Bayesian Methods.* Dordrecht, Holland: Kluwer Academic Publishing.

Sagan, C. (1994). Private statement from Sagan that he has used this assertion for many years, notably in the *Cosmos* TV series.

Sturrock, P. A. (1973). *Ap.* J., 182,569.

Sturrock, P. A. (1987). *Proposal for Evaluation of Remote-Viewing Information.* Presented at 6th S.S.E. Annual Meeting, Austin, Texas, May 29-30, 1987.

Sturrock, P. A. (1993). *Proposed Protocol for Evaluation of the Significance of UFO Reports.* Presented at 12th S.S.E. Annual Meeting, Santa Fe, New Mexico, June 24-26, 1993.

Sturrock, P. A. (1994). *Plasma Physics.* Cambridge University Press.

Ziman, J. (1978). *Reliable Knowledge. An Exploration of the Grounds for Belief in Science.* Cambridge University Press.