SURPRISE INDEX

INTRODUCTION

Perhaps the main function of a feeling of surprise is to make us reconsider the validity of our previous assumptions [8, p. 1131]. It can provoke us to change our subjective (personal) probabilities of various hypotheses and often to generate hypotheses that we had not previously entertained. These comments apply in ordinary life, in statistics, and even in mathematics. The topic of surprise is also of interest to some economists, psychologists, and philosophers [3,26]. Measures of surprise might also be of value (i) for constructing artificial music and other arts, where, to avoid monotony, some surprise is necessary, but not so much as to destroy the unity of the work; (ii) in a theory of humor where sudden changes in the frame of reference occur.

The first reasonable measure of surprise in terms of probability, apart from tail-area probabilities, was apparently proposed by Weaver [27,28]. Meanwhile, the economist Shackle [24,25] had proposed that the concept of potential surprise was fundamental in business decisions. This article surveys these ideas and later developments.

WEAVER’S SURPRISE INDEX

Although surprise is subjective, Weaver [27,28] suggested that an objective index of surprise could be defined that would measure the extent to which you (the subject) ought to be surprised. He emphasized that it would be an error to assume that an event of small probability should cause surprise: It has to be small compared with the probabilities of alternative outcomes. He considered an experiment having a discrete set of possible outcomes having probabilities $p_1, p_2, p_3, \ldots$. Then, if the $i$th of these outcomes occurs, his index of surprise is defined as $E_i[p_i]/p_i$, where $E_i$ denotes an expectation over the random variable $j$. This index is equal to $\rho/p_i$ where $\rho = \sum p_j^2$ is Gini’s index of homogeneity or the “repeat rate” in Turing’s later and self-explanatory terminology.

For the evaluation of Weaver’s surprise index for the Poisson and binomial distributions, see refs. 22 and 13, p. 562. For continuous distributions, one works with probability densities instead of probabilities, but the surprise index is then invariant only under linear transformations of the independent variable.

Any measure of surprise has to depend on the assumptions $H$ that we have before the observations are made, and a precise measure requires that $P(E|H)$ should be precise for each possible outcome $E$. In other words, either $H$ is a simple statistical hypothesis or else we need to be “sharp Bayesians” to have a sharp measure of surprise.

GENERALIZATIONS OF WEAVER’S INDEX: ENTROPY*

Weaver’s index is multiplicative if two independent experiments are combined into one, and there is a single-parameter generalization having the same property [7,8]. This is

$$\lambda_c = (\sum p_j^{c+1})^{1/c}/p_i, \quad c > 0, \quad (1)$$

where $c = 1$ gives Weaver’s index, while the limit as $c \to 0$ gives

$$\lambda_0 = p_i^{-1} \prod p_j^{p_i} = p_i^{-1} \exp \{E_j[\log p_j]\}. \quad (2)$$

The further generalization

$$p_i^{-1} \phi^{-1} \{E_j[\phi(p_j)]\},$$

where $\phi$ is a monotonic increasing function, is not multiplicative if $\phi$ is not a power or logarithm.

An additive index of surprise is $\Lambda_c = \log \lambda_c$, equal to the amount of information [6, p. 75] in the $i$th event minus the entropy [14]. The expression $\Lambda_c + \log p_i$ is sometimes called Rényi’s generalized entropy because of ref. 23, which, however, did not mention surprise indexes because he was unaware of refs. 7 and 8.

Because $E[\Lambda_0] = 0$, and because of its close relationship to entropy and information, it seems that $\Lambda_0$ is the most natural of these
additive surprise indexes, and $\lambda_0$ the most natural multiplicative one. Negative values of $\Lambda_0$ correspond, so to speak, to nondescript outcomes or “antisurprise” [15]. Bartlett [2] discussed $\Lambda_0$ in relation to “the significance of odd bits of information,” but without explicit reference to surprise indexes.

For the $k$-dimensional multivariate normal distribution $N(\mu, C)$, we have [8]

$$\Lambda_c = \frac{1}{2} (x - \mu)' C^{-1} (x - \mu) - \frac{k}{2} \log (c + 1), \quad c > 0,$$

(3)

$$\Lambda_0 = \frac{1}{2} (x - \mu)' C^{-1} (x - \mu) - \frac{1}{2} k$$

(4)

where $x$ is the observed value of the random vector and $D^2$ is the Mahalanobis (squared) distance between $x$ and $\mu$. Note that $D^2$ has a chi-squared distribution with $k$ degrees of freedom.

**DEPENDENCE ON THE CATEGORIZATION**

Although the probabilities of all hands of 13 cards in the game of bridge have equal probabilities, some hands, such as all 13 spades, are of special human interest and for that reason would be surprising. Many other hands are also interesting to various degrees. As emphasized in ref. 8, the surprise indexes defined so far depend very much on the way that you have categorized the outcomes. This fact largely undermines the objectivity (impersonal character) of the surprise indexes in many circumstances. Previous information and hypotheses will also change the degree of surprise because they change your subjective probabilities. For example, an all-spades hand is much less surprising if, before the cards were dealt, you had noticed the other players exploding with irrepressible mirth.

**SHACKLE’S POTENTIAL SURPRISE**

Unaware of Weaver’s note, Shackle [24,25] used the concept of potential surprise, instead of degrees of belief, to attack the question of how people, especially entrepreneurs, make decisions. He considered that the “interest-ness” of an imagined outcome was a function of its desirability and of its potential surprise, and that people, when deciding on an action, usually concentrate on two “focus outcomes” of maximum interest-ness, one desirable and the other undesirable. A feeling of surprise is an emotion, whereas a judgment of subjective probability is more intellectual, so perhaps decisions, especially emotional ones, are often made somewhat along Shackelian lines.

Good [7] argued that, since surprise could be given various meanings in terms of subjective probability, it should be possible to use judgments of probabilities to sharpen judgments of potential surprise and vice versa. In this way, your entire body of judgments might be improved. Krelle [20] argued that there is a one-to-one relationship between degree of potential surprise and subjective probability.

**A WEAKNESS OF ADDITIVE SURPRISE INDEXES: TAIL-AREA PROBABILITIES**

The example of the multivariate normal reveals a weakness in all these surprise indexes, namely that they can more easily exceed a given value when the dimensionality is increased. This weakness is a simple consequence of the additivity property alone. It may be better to treat a surprise index, as defined so far, as a statistic whose tail-area probability $P$, or better $P^{-1}$, is used as the revised index of surprise. This would be a kind of surprise-Fisher compromise. It would be consistent with the treatment of “the significance of odd bits of information” by Bartlett [2].

It would be possible to use, as a measure of surprise, $P^{-1}$, where $P$ is the tail-area probability of any statistic $S$ used for testing our prior beliefs. To take $S$ as $\Lambda_0$ is equivalent to defining a surprise index as the sum of the probabilities of all events whose probabilities do not exceed $P(E|H)$. This definition still depends, for discrete variables, on how the events are categorized, and, for continuous variables, is not invariant under nonlinear transformations of the independent variable(s). The possibility of attaining
invariance will be discussed later in this article.

A PRINCIPLE OF LEAST SURPRISE

Good [9,16] suggested, but did not strongly advocate, the possibility, after an observation is made, of selecting a hypothesis \( H \) (or estimating a parameter, which is logically the same thing), by a principle of least surprise, and that, if this is done, the prior (initial) probability \( P(H) \) should also be taken into account. A special case of this suggestion, without allowing for complexity, was proposed by Barndorff-Nielsen [1], who was unaware of ref. 9.

If the index \( \Lambda_0 \) is used, the expressions to be minimized are, respectively,

\[
-\log P(E_i|H) + \sum_j P(E_j) \log P(E_j|H) \quad (5)
\]

and

\[
-\log P(E_i;H) + \sum_j P(E_j) \log P(E_j|H), \quad (6)
\]

depending on whether \( P(H) \) is or is not taken into account. Here \( E_1, E_2, \ldots \) denote the possible outcomes of an observation and \( E_i \) is the one that occurred. Non-Bayesians might prefer to minimize (5) rather than (6), though the two procedures are equivalent if all the hypotheses under consideration have equal prior probabilities.

Fortunately, simple hypotheses often have higher prior probabilities than complicated ones, so that the capacity of surprise leads to the discovery of new truths [8, p. 1131]. As often happens in the application of significance tests, a surprising outcome can cause us to look for new models or hypotheses. This is true whether the assumed model is non-Bayesian or Bayesian.

AN INVARIANT INDEX RELATED TO COMPLEXITY

The concept of surprise is closely connected with those of complexity and coincidences, but the surprise indexes mentioned before do not explicitly allow for complexity. To understand the connections, let us consider as an example the true mathematical assertion \( E \), that

\[
|163 - [\pi^{-1} \log_e(640320^3 + 744)]^2| < 10^{-32}. \quad (7)
\]

For this example, it seems difficult or impossible to apply the surprise indexes mentioned before.

To decide whether we should be surprised by (7), define a proposition \( F \) as the logical disjunction of all propositions \( E_{a,b} \) of the form that the difference between \( [\pi^{-1} \log_e(a^3 + b)]^2 \) and the closest integer to it is less than \( 10^{-32} \), where \( a \) and \( b \) are positive integers and \( a < 10^4, b < 10^5 \). Given a naive (but by no means stupid) state of mathematical knowledge, \( H_0 \), the probability of \( E_{640320,744} \) is \( 2 \times 10^{-32} \). (This is a "dynamic probability," whereas the logical probability is 1; see AXIOMS OF PROBABILITY.) But (7) is much less surprising than if two genuine independent randomizing devices, when started, both produced the same sequence of 32 decimal digits. This is largely because we must "pay" for the complexity of \( F \). [For attempts to measure complexity, see, for example, refs. 17, 19, (pp. 155 and 235), and 4.] To estimate an upper bound to a measure of the complexity, we may generously allow one "decimal unit of surprise" for each of \( \pi, -1, \log_e, a, 3, +, b, 2, 10, 6, 10, \) and 3 and at most four more for the remaining syntactic structure of the statement, and thus count the complexity of \( F \) as no more than 16 decimal units. We are still left with at least another 15.7 decimal units of surprise. So (7) is much too surprising to be a coincidence and it must have a non-number-crunching "explanation," whether or not any one knows it. The explanation is in fact known: It involves the theory of the elliptic modular function [29, p. 461; 10]. When a mathematician confidently conjectures a theorem, it is perhaps because he believes that the evidence would be too surprising if the theorem were known to be false.

This example suggests that a reasonable index of surprise, conditional on previous assumptions \( H_0 \), is

\[
S(E|H_0) = -\log_{10} P(E|H_0) - \chi(E|H_0), \quad (8)
\]
where $\chi^2(E|H_0)$ is an additive measure, in decimal units, of the complexity of the part of $E$ that goes beyond what is known to follow from $H_0$. The formula (8) differs from that for $\Lambda_0$ in that the entropy term in $\Lambda_0$ is replaced by a complexity term. Note that, should $E$ contain real parameters measured to unnecessarily many decimal places, the value of $S(E|H_0)$ would be unchanged. This is a necessary invariance property for any satisfactory index of surprise. Further, if $H_0$ is a necessary invariance property for any value of $H$ to unnecessarily many decimal places, the should replaced by a complexity term. Note that, by-two contingency tables example. Pearson [21] collected 12448 two-by-two tables would be no surprise, as is appropriate. Both terms of (8) would vanish, that is, there were replaced by a full explanation of hypothesis of least surprise, if its utterly negligible region of four or five decimal units. A hypothesis had not been concerned with the values of chi-squared. There are several relevant possible hypotheses, so we omit details, but, at first sight at least, a measure of surprise is in the region of four or five decimal units. A hypothesis of least surprise, if its utterly negligible initial probability is not taken into account, is that Pearson thought the expected value of chi-squared was 1 and that he cheated, without even mentioning chi-squared in his paper! Never would anyone go to so much trouble with so little purpose. This example illustrates that, when we are convinced that the surprise in an event can be removed only by making an entirely unreasonable hypothesis, then the event is called a pure coincidence [11, p. 169; 17, p. 146]. The example shows that a principle of least surprise that does not allow for prior probabilities of hypotheses is a seriously incomplete recipe for selecting a hypothesis.

REFERENCES

2. Bartlett, M. S. (1952). Biometrika, 39, 328–337. (Uses $\Lambda_0$ to measure the significance of odd bits of information as defined in ref. 9, without reference to surprise.)
4. Cover, T. M. (1973). Ann. Statist., 1, 862–871. (Suggests that the complexity of $H$ could be measured by the number of hypotheses at least as “simple” as $H$. The logarithm of this number is perhaps preferable.)
5. Fisher, R. A. (1926). Eugenics Rev., 18, 32–33. (Found that the average chi-squared of 12448 two-by-two tables was remarkably close to 1. See ref. 21.)
12. Good, I. J. (1982a). J. Amer. Statist. Ass., 77, 342–344. (Emphasizes that if an observation is surprising given both the null and the non-null hypothesis, then a search for other hypotheses is sensible.)
surprise index for binomials is expressible in terms of Legendre polynomials.)

14. Good, I. J. (1983a). Behav. Brain Sci., 6, 70. (Points out that the basic formula in the article under discussion is that for λ₀.)


18. Good, I. J. (1984). J. Statist. Comp. Simul., 20, 294–299. (Suggests that, given some observation, the maximum explicativity is a better criterion than least surprise, when choosing a hypothesis. Dynamic probability must be used if the hypotheses are unconstrained.)


21. Pearson, E. S. (1925). Biometrika, 17, 388–442. (Discusses his collection of 12448 two-by-two contingency tables that were later used in ref. 5.)


See also BAYESIAN INFERENCE; BELIEF, DEGREES OF; COMBINATION OF DATA; FIDUCIAL INFERENCE; INFERENCE, STATISTICAL—I; INFERENCE, STATISTICAL—II; PRIOR PROBABILITIES; SUBJECTIVE PROBABILITIES; AND TWO-BY-TWO (2 × 2) TABLES.

I. J. GOOD