THE SURPRISE INDEX FOR THE MULTIVARIATE NORMAL DISTRIBUTION

By I. J. Good

1. The surprise index and its generalisations. Let $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \ldots$ be a natural classification into a finite or countably infinite number of possible mutally exclusive and exhaustive results of some experiment or observation, and let $P(\mathbf{E}_i \mid H) = p_i (i = 1, 2, 3, \cdots)$, where H is a simple statistical hypothesis. Then the surprise index (Weaver [7]) associated with the result $\mathbf{E}_i$ is

$$\lambda_1 = \frac{E_j(p_j \mid H)}{p_i} = \frac{\sum_j p_j^2}{p_i}. \tag{1}$$

If the experiment consists in the measurement of a continuous vector or scalar variable with a differentiable distribution function, we define

$$\lambda_1 = \frac{E(p^* \mid H)}{p}, \tag{2}$$

where p^* is the random variable that is the probability density of the original random variable, and p is a realisation of p^*.

For practical purposes, (2) is almost the same definition as (1). For example, a continuous scalar variable is usually measured to some fixed number, n, of decimal places, and the natural classification of the possible results of the experiment is into intervals of length 10^{-n} of values of the variable. If we then use definition (1) and let n tend to infinity, we get definition (2). For experiments with results that are real variables having distributions that are partly discrete (atomic) and partly continuous (differentiable), it is not immediately obvious what definition should be used. Something more will be said about this later.

The surprise index is open to two criticisms:

(I) It is changed when the results of an experiment are lumped together in a new way, in the discrete case, or when there is a change of mathematically independent variable in the continuous case.

(II) The numerator in (1) or (2) is somewhat arbitrary.

We shall now discuss these two criticisms.

As an example of (I), suppose that an "unbiased coin" is spun twenty times.[1] There is an obvious classification of the possible results of the experiment into 2^{20} categories. But, with this classification,

$$HTHTHTHTHTHTHTHTHTHT \tag{3}$$

Received September 22, 1955.

[1] By putting the description "unbiased coin" in quotation marks, we intend to imply that a certain self-explanatory simple statistical hypothesis is to be taken for granted, and the probabilities of the possible results are the tautological ones usually associated with this idealised experiment.

has the same surprise index as

(4) *HTTHTHHHTTHHHHTTTHTH.*

In practice, (3) would be more surprising than (4), at any rate if neither of them had been written down in advance of the experiment. This is partly because (3) is simpler. (The reader should avoid being confused by the two meanings of the word "simple." We use the word in its technical sense only in the phrase "simple statistical hypothesis," while in "simple hypothesis," the word has its ordinary non-technical meaning.)

If we imagine that the 2^{20} possible results are classified into groups of roughly equal simplicity, (3) would belong to a small group, whereas (4) would belong to a large group. If we regard all the results in one group as a single possible result, it follows that (3) would, after all, have a higher surprise index than (4). Thus the vagueness of definition (1) is seen to arise from the difficulty of measuring simplicity. (With regard to the regrouping of results, see Bartlett [1], page 231.)

The connection between surprise and simplicity can be defended by the following argument.

Perhaps the main biological function of surprise is to jar us into reconsidering the validity of some hypothesis that we had previously accepted. Hence, we tend to be surprised when evidence is received against such a hypothesis, i.e., when the result of an observation has much greater probability when given some other, not entirely untenable, hypothesis. But in the process of being surprised, we often do not have time to estimate the initial probability of the rival hypothesis; instead, we tend to notice whether the rival hypothesis is very simple. More formally, we are surprised if $\mathbf{E}$ occurs when the likelihood ratio $P(\mathbf{E} \mid H') / P(\mathbf{E} \mid H)$ is large, where H was previously believed and H' is very simple.

Fortunately, simple hypotheses often have higher initial credibilities than complicated ones, so that the capacity of surprise leads to the discovery of new truths.

In the above example, a hypothesis that would explain (3) would be that the coin always, or very often, rotates by the same (odd) number of half-revolutions.

Since no one has yet thought of a satisfactory measure of simplicity, it seems unlikely that a really satisfactory measure of surprise can be given. For an experiment whose result is naturally expressed as a single integer, the difficulty does not seem to matter greatly. It is true that we may be temporarily surprised because the integer has striking properties, like those of 10,000 or 22,222, but we are often able to discount this sort of surprise as being due to a "mere coincidence" and as being dependent on the irrelevancy that we use radix 10.

Obvious examples of experiments whose results are integers are those giving rise to binomial and Poisson distributions. For these, λ_1, has been evaluated by Redheffer [6]: for the binomial distribution, λ_1 is expressible in terms of the

sums of the squares of the binomial terms (not coefficients) and therefore in terms of Legendre polynomials. Outside the range of existing tables, the Legendre polynomials that occur here may be conveniently computed with the help of a formula given by Good [4].

We now consider criticism (II). A generalisation of the surprise index, with a more general numerator, has been briefly discussed by Good [3]. Let

$$\lambda_0 = \frac{[E(p^{*u})]^{1/u}}{p} \qquad (u > 0),$$

$$\lambda_0 = \exp\{E(\log p^*) - \log p\} = \text{G.E.}(p^*)/p,$$

(where G.E. means "geometric expectation"), and let

$$\Lambda_u = \log \lambda_u \qquad (u \geqq 0).$$

We may call Λ_u a "logarithmic surprise index." It can be seen at once that $\lambda_u (u \geqq 0)$ is multiplicative, whereas Λ_u is additive, if the results of several statistically independent experiments are combined into a single experiment. Weaver [7] did not allow his surprise index to be less than 1, but it is necessary to do so in order to achieve multiplicativity. A negative logarithmic surprise index corresponds to an event that "was only to be expected."

Of the continuous infinity of surprise indexes, the most natural ones seem to be λ_1 and λ_0, or, equivalently, Λ_1 and Λ_0. Bartlett [1] discussed Λ_0, but not in relation to Weaver's suggestion. We shall argue below that λ_0 (or Λ_0) is rather better than λ_1, at any rate for multivariate normal distributions. For univariate normal distributions, there is little difference between Λ_0 and Λ_1.

Before going on to this, we shall digress for a moment in order to discuss (i) distributions that are partly discrete (as promised earlier) and (ii) tail-area probabilities.

2. Partly discrete distributions. The above reference to multiplicativity suggests a possible definition of λ_u for a univariate distribution that is partly discrete and partly continuous. We can first classify the possible results into "atomic" on the one hand and "non-atomic" on the other. This is a two-category (discrete) classification for which $\lambda_u^{(1)}$ may be defined as λ_u was before. Then, if the observed value of the random variable is atomic (or non-atomic), we can compute a conditional $\lambda_u^{(2)}$, i.e., conditional on the information that the variable is an atomic one (or a non-atomic one). Finally, we can define $\lambda_u = \lambda_u^{(1)}\lambda_u^{(2)}$.

3. Tail-area probabilities. The so-much-or-more method in statistics is the usual method in which the result of an experiment or observation is summarised by means of a tail-area probability

$$P(x^* > x), \qquad P(x^* \geqq x), \qquad \text{or} \quad P(x^* > x) + \tfrac{1}{2}P(x^* = x),$$

where x^* is a real random variable and x is a real number. This method is most satisfying when x^* is the likelihood of an experiment (given a null hypothesis), but in this case the tail-area probability is often difficult to evaluate numerically.

Moreover, there are again logical difficulties for distributions that are partly discrete and partly continuous.

The reciprocal of a tail-area probability is often not more than about 10 times the Bayes factor against the null hypothesis, calculated in accordance with some reasonable assumptions about the initial distributions and probabilities (See Good [2], page 94.) When the ratio is greater than about 10, there is likely to be some argument about which is the better statistic. This difficulty can easily arise for bimodal distributions.

Jeffreys [5], page 316, says "What the use of P (a tail-area probability) implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred." In other words, a tail-area probability consists in the probability of an experimental result artificially added to the probabilities of results that did not occur, or, if not artificially, at any rate with incomplete logical justification.

λ_u, for any u, as a final summary of an experiment or observation, overcomes Jeffreys' criticism of a tail-area probability, although it may still be unsatisfactory as compared with upper and lower bounds for a Bayes factor when we are prepared to assume enough about the non-null hypothesis. For a distribution with density such as

$$\frac{1}{\sqrt{8\pi}}\left(e^{-(x+4)^2/2}+e^{-(x-4)^2/2}\right),$$

λ_u is apt to be a much better summary of the experiment than a tail-area probability would be. But it can be argued that better still would be the tail-area probability associated with the value of λ_u. This would come to the same thing as the use of the distribution of the likelihood or the likelihood density. (The possibility of using Weaver's surprise index, λ, as a substitute for the use of tail-area probabilities was suggested in conversation by Mr. G. C. Wall.)

4. λ_u for multivariate normal distributions. For multivariate normal distributions, $P(p^* < p)$, the distribution of the likelihood density, does not seem to be expressible in elementary terms. It is therefore perhaps more worth while to compute λ_u for the multivariate normal distribution than for the Poisson and binomial distributions.

A k-dimensional multivariate normal distribution has a density function of the form

$$p = \frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}k}} \exp\left\{-\tfrac{1}{2}\sum_{i,j}^{1,2,\cdots,k} A_{ij}(x_i-a_i)(x_j-a_j)\right\},$$

where $|A| = \det\{A_{ij}\}$. (See, for example, Wilks [8], p. 65.)

Now, it is easily seen that for any k-dimensional probability density, the generalised surprise indexes λ_u, $\Lambda_u (u \geqq 0)$ are invariant under all non-singular linear transformations.[2] This observation follows from the fact that the Jacobian

[2] The method of this paragraph is due to the referee; my own method was clumsier.

of such a transformation is constant and non-zero. Therefore, for non-degenerate k-dimensional multivariate normal distributions, there is no real loss of generality in taking $A = I$ (the identity matrix) and $a_1 = a_2 = \cdots = a_k = 0$. For this standardised distribution, we can use the multiplicative property of λ_u for probabilistically independent experiments, together with a simple univariate integration, to evaluate λ_u. Then, transforming back to the general non-singular distribution, we get

$$\lambda_u = \frac{1}{(u+1)^{k/2u}} \exp\{\tfrac{1}{2}\sum_{i,j} A_{ij}(x_i - a_i)(x_j - a_j)\} \qquad (u > 0),$$

$$\Lambda_u = \tfrac{1}{2}\sum_{i,j} A_{ij}(x_i - a_i)(x_j - a_j) - \frac{k}{2u}\log(u+1) \qquad (u > 0),$$

$$\Lambda_0 = \tfrac{1}{2}\{\sum_{i,j} A_{ij}(x_i - a_i)(x_j - a_j) - k\}.$$

It may be observed that Λ_u (and therefore λ_u), regarded as a function of u, is continuous to the right at $u = 0$. By writing $u = e^v - 1$, we see at once that Λ_u is a strictly increasing function of u. When $u \to \infty$, Λ_u tends to

$$\Lambda_\infty = \tfrac{1}{2}\sum_{i,j} A_{ij}(x_i - a_i)(x_j - a_j).$$

From this expression it is clear that Λ_∞ is the logarithm of the likelihood ratio in the sense of Wilks [8] for testing the hypothesis of our multivariate normal distribution "within" the more general class of multivariate normal distributions that have the same matrix $\{A_{ij}\}$, or, what comes to the same thing, the same covariance matrix.

It is known (see, for example, Wilks [8], page 104) that $2\Lambda_\infty$ has precisely a chi-squared (gamma-variate) distribution with k degrees of freedom. Since

$$\Lambda_u = \Lambda_\infty - \frac{k}{2u}\log(u+1),$$

we can obtain the exact tail-area probability corresponding to any observed value of Λ_u. But we may also develop an intuitive appreciation of Λ_u (or λ_u) in itself, for some fixed value of u. In order to decide which is the most natural value of u to take, we note that $E(\Lambda_0) = 0$. (This is obvious from the definition of Λ_0 and also from the fact that $2\Lambda_0 + k$ has a chi-squared distribution with k degrees of freedom.) It seems natural to demand that the expected log-surprise should be zero before an experiment is performed. It is not equally natural to insist that $E(\lambda_u) = 1$, or that $E(\lambda_u^{-1}) = 1$ (which gives $u = 1$), since λ_u and λ_u^{-1} have very skew distributions. For very skew distributions, expected values are more artificial than for ordinary distributions such as the chi-squared. For one thing, the median is a long way from the expected value for very skew distributions.

We conclude, then, that for the k-dimensional multivariate normal distribution,

Λ_0 and λ_0 seem more natural measures of surprise than Λ_1 and λ_1, whereas other values of u do not seem to have anything special to commend them. There is little difference between Λ_0 and Λ_1 when k is small.

For $k = 1$, we have

$$\lambda_0 = e^{s^2/2} / \sqrt{e}, \qquad \lambda_1 = e^{s^2/2} / \sqrt{2},$$

where s is the "sigma-age" of an observation; i.e., the deviation from the mean divided by the standard deviation. Some numerical values of λ_0 and λ_1 are given in the following table, together with the reciprocals of the corresponding two-tailed tail-area probabilities, $P(s)$.

s	0	1	2	3	4	5
$1/P(s)$	1	3.1	22	370	16000	1740000
λ_0	0.61	1	4.5	54.6	1800	160000
λ_1	0.71	1.17	5.2	64	2100	187000

If we have a sample of several independent observations (k-dimensional vectors) from our multivariate normal distribution, we can compute λ_0 for the whole sample by multiplying together the separate λ_0's. This method may be regarded as an alternative to Hotelling's generalised "Student" test. (See, for example, Wilks [8], Section 11.4, where further references are given.)

REFERENCES

[1] M. S. Bartlett, "The statistical significance of odd bits of information," *Biometrika*, Vol. 39 (1952), pp. 328–337.

[2] I. J. Good, *Probability and the Weighing of Evidence*, Charles Griffin, London, 1950.

[3] I. J. Good, "The appropriate mathematical tools for describing and measuring uncertainty," *Uncertainty and Business Decisions*, C. F. Carter, G. P. Meredith, and G. L. S. Sheckle (eds.), University Press, Liverpool, 1954.

[4] I. J. Good, "A new finite series for Legendre polynomials," *Proc. Cam. Phil. Soc.*, Vol. 51 (1955), pp. 385–388.

[5] H. Jeffreys, *Theory of Probability*, Oxford University Press, 1939.

[6] R. M. Redheffer, "A note on the surprise index," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 128–130.

[7] W. Weaver, "Probability, rarity, interest and surprise," *Scientific Monthly*, Vol. 67 (1948), pp. 390–392.

[8] S. S. Wilks, *Mathematical Statistics*, Princeton University Press, 1946.