DOCID: 3838681

A List of Properties of Bayes-Turing Factors

BY I. J. GOOD

Unclassified

Just as Bayes deserves and gets credit for noticing an interesting but mathematically trivial consequence of the product axiom of probability, so Jeffreys and Turing deserve credit for noticing a further mathematically trivial but interesting consequence of Bayes' theorem. This consequence is expressible in terms of what is often called a Bayes factor, although it would be as just to ascribe Bayes' theorem to Fermat, who was possibly the first person to enunciate the product axiom.

In this paper we list some of the properties of the Bayes factor, without proofs. Most of the proofs are simple, available elsewhere, and omitted. No examples will be given: a list is a list is a list. But a few historical notes are included. See also Reference [7].

Let E denote an experimental result or evidence, let H denote a hypothesis, and let $\overline{H}$ denote its negation. We use the symbol $\cup$ to mean "or", so that, for example, $A \cup B$ means *either A or B or both*, and a period for "and." When H and H' are hypotheses and $H \cup H'$ is taken for granted, then $\overline{H}$ is another notation for H'. The probability of E given H is denoted by $P(E|H)$, where the vertical stroke is pronounced "given." If p is a probability, the corresponding *odds* are defined as $p/(1 - p)$. It is a simple consequence of the product axiom of probability that

$$\frac{O(H|E)}{O(H)} = \frac{P(E|H)}{P(E|H)}, \tag{1}$$

where $O(H)$ are the odds of H.

Bayes' theorem, in the form stated by Jeffreys in his *Theory of Probability*, asserts that

$$P(H|E) \text{ is proportional to } P(H)P(E|H), \tag{2}$$

when there are several hypotheses. [Proof: $P(H|E) = P(H)P(E|H)/P(E)$, end of proof.] To obtain (1), replace H by $\overline{H}$ in (2) and divide. Jeffreys noticed (2) in 1935, but in his book he invariably takes $O(H) = 1$, and so obscures the issue.

Turing fortunately did not know in 1940 that $P(E|H)$ is called the likelihood of H given E, since he otherwise might not have named

Approved for Release by NSA on Appeal on 03-9-2011, FOIA Case # 58820

$O(H|E)/O(H)$ *the factor in favour of H provided by E*, which is a phrase of immediate intuitive appeal. A pronounceable notation for it is $F(H{:}E)$, where the colon is pronounced *provided by*. It is the factor by which the initial odds of H are multiplied, in the light of the evidence, in order to get the final odds. Its logarithm is called the "log-factor" or the "weight of evidence," $W(H{:}E)$, in favour of H provided by E. When the base of the logarithms is 10, Turing called the unit in which the log-factor is measured a *ban* or ten decibans, by analogy with "decibels." If we wish to make it clear that we are assuming $H_1 \cup H_2$, a convenient notation is $F(H_1{:}E|H_1 \cup H_2) = F(H_1/H_2{:}E)$, pronounced *the factor in favour of* H_1 *as against* H_2, *provided by* E. The notations

$$W(H_1/H_2{:}E),\ F(H_1/H_2{:}E|H_3),\ W(H_1/H_2{:}E|H_3),$$

and their pronunciations, are now self-evident.

When H and $\overline{H}$ are "simple statistical hypotheses," equation (1) states that the factor in favour of H is equal to the likelihood ratio. But in more general circumstances the "likelihood ratio" means a ratio of maximum likelihoods, and then the Bayes factor is *not* equal to a likelihood ratio. Even when equality holds, the expression "factor in favour of a hypothesis" puts the emphasis on the other side of the equation and so is not semantically identical with "likelihood ratio."

Theorem 1.

$$W(H{:}E) = I(H{:}E) - I(\overline{H}{:}E)\ ,$$

where $I(H{:}E)$ is the amount of information concerning H provided by E, i.e.,

$$I(H{:}E) = \log \frac{P(H \cdot E)}{P(H)P(E)}\ .$$

Theorem 2. If there is more than one piece of evidence, and if these pieces of evidence are statistically independent given H and also given $\overline{H}$, then the factors are multiplicative and the log-factors are additive.

Theorem 3. If an experiment has various possible experimental results, $E_1, E_2, \ldots$, then the expected log-factor in favour of H, when H is true, is positive, and, when H is false, is negative. More shortly: an expected log-factor for a true hypothesis is positive. (The expectation is zero only if the experiment is irrelevant to H.)

Theorem 4. If H_1 and H_2 are mutually exclusive, $H = H_1 \cup H_2$, and $P(E|H_2) = P(E|\overline{H})$, then

$$F(H:E) = pf + 1 - p ,$$

where $p = P(H_1|H)$, $f = F(H:E|H_1 \cup \overline{H})$.

Theorem 5. (The "weighted average of factors")

(i) Let $H_1, H_2, H_3, \ldots$ be mutually exclusive, and let $H = H_1 \cup H_2 \cup H_3 \cup \ldots$. Then

$$F(H:E) = \Sigma p_i f_i ,$$

where

$$p_i = P(H_i|H), f_i = F(H:E|H_i \cup \overline{H}) .$$

(ii) Let $H_1, H_2, H_3, \ldots$ be mutually exclusive and let $H' = H_1 \cup H_2 \cup H_3 \cup \ldots$. Then

$$F(H:E|H') = \Sigma_i P(H_i|H \cdot H') \frac{P(E|H \cdot H_i)}{P(E|\overline{H} \cdot H')} ,$$

and, if $P(E|\overline{H} \cdot H_i)$ is mathematically independent of i,

$$F(H:E|H'] = \Sigma_i P(H_i|H \cdot H') F(H:E|H_i) .$$

(iii) Let $H_1, H_2, H_3, \ldots$ be mutually exclusive and exhaustive. Then

$$F(H:E) = \Sigma_i P(H_i|H) P(E|H \cdot H_i)/P(E|\overline{H}) ,$$

and, if $P(E|H \cdot H_i)$ is mathematically indepenent of i,

$$F(H:E) = \Sigma_i P(H_i|H) F(H:E|H_i) .$$

Theorem 6. If evidence arrives in a large number of small independent pieces, then the total log-factor will have approximately a normal distribution if H is true, and also if H is false. (This can be rigorously stated by reference to any rigorous statement of the Central Limit Theorem.)

Theorem 7. Let $x = W(H:E)$. Then

$$W(H:E) = W\{H:W(H:E)\} = W(H:x) ,$$

so that

$$P(x|H) = e^x P(x|\overline{H}) .$$

Theorem 8. Under the conditions of Theorem 5,

$$\mathbf{E}(W(H:E)|H) \approx -\mathbf{E}(W(H:E)|\overline{H}) ,$$

i.e., the expected log-factor in favour of H if H is true is approximately equal and opposite to the expected log-factor in favour of H if H is false.

Theorem 9. Also under the conditions of Theorem 5,

$$\begin{aligned}\mathrm{var}(W(H{:}E)|H) &\approx \mathrm{var}(W(H{:}E)|\overline{H}) \\ &\approx 2\mathbf{E}(W(H{:}E)|H)\ ,\end{aligned}$$

if natural logarithms are used.

Corollary.

$$S.D.\ (W(H{:}E)|H) \approx 3\sqrt{\mathbf{E}(W(H{:}E)|H)}$$

if the unit is the deciban.

Theorem 10. ("False alarm probabilities.") Suppose that the distribution of $x = W(H{:}E)$, given H, is approximately normal in the neighbourhood of $x_o = \mathbf{E}(W(H{:}E)|H)$, in "natural bans" (logarithms to base e). Let a "threshold" t be selected near x_o. Then the "false alarm" probability

$$P(x > t|\overline{H}) \approx e^{(1/2)\sigma^2 - x_o}\ \psi\left(\sigma + \frac{t - x_o}{\sigma}\right),$$

where

$$\psi(u) = \frac{1}{\sqrt{2\pi}}\int_u^\infty e^{(-1/2)z^2}\,dz\ ,$$

$$\sigma^2 = \mathrm{var}(x|H)\ .$$

Theorem 11. Let an experiment be performed, of various possible results, such as E, and let

$$F(H{:}E) = f\ .$$

Then

$$\mathbf{E}(f|\overline{H}) = 1\ .$$

Theorem 12. In the notation of Theorem 11,

$$\mathbf{E}(f^n|H = \mathbf{E}(f^{n+1}|\overline{H})\ ,$$

for any real number n; for example,

$$\mathbf{E}(f^{-1/2}|H) = \mathbf{E}(f^{1/2}|\overline{H})\ .$$

More generally, for any function $\psi(f)$,

$$\mathbf{E}\{\psi(f)|H\} = \mathbf{E}\ \{f\psi(f)|\overline{H}\}\ .$$

Theorem 13.

$$P(f \leq g|H) \leq g\ .$$

Theorem 14. Let an experiment, E, have only two possible outcomes E_1 and E_2, where $P(E_1|H) > P(E_1|\overline{H})$. Then

DOCID: 3838681

$$\mathbf{E}\{W(H:E)|H\} < |\mathbf{E}\{W(H:E)|\bar{H}\}|$$

if and only if $P(E_1|H) > P(E_2|\bar{H})$.

NOTES CONCERNING THE THEOREMS

Theorem 1. This is mentioned because it shows perhaps the simplest connection between Turing's Theory of evidence and Shannon's theory of information. The advantage of using the term "evidence" or "weight of evidence," rather than "information," for the log-factor is that the log-factor was in use long before Shannon's work. It is historically misleading to call the log-factor "information," since it gives the false impression that it was derived from Shannon's work.

Theorem 2. In view of this theorem, Turing pointed out the possibility of a sequential use of log-factors, in which one can accept H when the cumulative log-factor reaches some threshold, or reject it if the cumulative log-factor falls below some other threshold. He also pointed out that one could estimate in advance the number of trials required in order to achieve some threshold, but he did not work out any of the relevant random-walk theory. A. Wald carried this part of the theory much further when he independently introduced the sequential log-factor method some years later. (Ref. [1].)

Theorem 3 shows that it can never be wrong in expectation to look at new evidence, if the cost of doing so be ignored. This is intuitively obvious, and so corroborates that we are using a sensible definition of weight of evidence. (See also Ref. [3].)

Theorems 2, 3, and 4 are due to Turing. Theorem 5 is my generalization. (5(i) was stated on p. 68 of Ref. [2]. Chapter VI of Ref. [2] mentions several of the results of the present paper.)

Theorem 6. Turing's Fellowship thesis at King's College, Cambridge, was concerned with the Central Limit Theorem, but he did not publish it because he found that the work had already been done by Feller in 1935.

Theorems 6 to 9 are due to Turing. (The corollary to Theorem 9 is my trivial transformation of it.) (See p. 73 of Ref. [2], and pp. 129–130 of Ref. [4].) The generalization, Theorem 10, was noticed independently by R. B. Dawson and myself. (See Ref. [4], which contains further results. See also Ref. [5] and [6].) Theorem 11 is Turing's and Theorem 12 is a generalization of mine.

Theorem 13 was first drawn to my attention by G. C. Wall in 1949. Like some other results in this paper, this one is probably in Ref. [1]. Theorem 14 was noticed by me in 1948.

REFERENCES

[1] Abraham Wald, *Sequential Analysis*, John Wiley & Sons, New York, Chapman and Hall, London, 1947.

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

[3] ————, "Weight of evidence, corroboration, explanatory power, information, and the utility of experiments," *J. Roy. Statist. Soc. B, 22*, 1960, pp. 319–331.

[4] ————, "Weight of evidence, causality, and false-alarm probabilities," *Fourth London Symp. on Information Theory*, Butterworths, London, 1961, pp. 125–136.

[5] ————, "Weighted covariance for estimating the direction of a Gaussian source," *Time Series Analysis: Proc. of the Symp. held at Brown University, June 11–14, 1962*, John Wiley & Sons, New York and London, 1963, pp. 447–470.

[6] W. W. Peterson, T. G. Birdsall, and W. C. Fox, "The theory of signal detectability," *Trans. I.R.E. PGIT* Vol. 4 (1954), pp. 171–212. [This is *not* IT–4, 1958!]

[7] R. P. Dilworth, "Bayes Factors," *IDA–CRD Expository Report No. 1*, October, 1962.