Biometrika (1979), **66**, 2, *pp.* 393–6
Printed in Great Britain

Studies in the History of Probability and Statistics. XXXVII A. M. Turing's statistical work in World War II

By I. J. GOOD

Department of Statistics, Virginia Polytechnic Institute & State University, Blacksburg

Summary

An account is given of A. M. Turing's unpublished contributions to statistics during 1941 or 1940.

Some key words: Bayes factors; Cryptology; Decibans; Diversity; Empirical Bayes; History of statistics; Information; Repeat rate; Sequential analysis; Weight of evidence; World War II.

1. Preamble

Alan Mathison Turing (1912–54) is best known for the concept of the Turing Machine (Turing, 1936–7). He introduced this concept in his proof that no finite decision procedure can solve all mathematical problems. Owing to a security curtain that lifted only a few years ago, it is less well known that he made important contributions to cryptanalysis during World War II. I was familiar with much of this work because of being his main statistical assistant in 1941. During the war, part of his work related to electromagnetic and electronic machinery, but I shall deal here only with his statistical ideas. All of these date from 1941 or 1940. These statistical ideas are not treated in the biography of Turing by his mother (Turing, 1959).

2. Bayes factors

In practical affairs and in philosophy it is useful to introduce intuitively appealing terminology. Turing introduced the expression '(Bayes) factor in favour of a hypothesis', without the qualification 'Bayes'. The (Bayes) factor in favour of a hypothesis H, provided by evidence E, is $O(H \mid E)/O(H)$, the factor by which the initial odds of H must be multiplied to get the final odds. It is an easy but important theorem that the Bayes factor is equal to $\mathrm{pr}(E \mid H)/\mathrm{pr}(E \mid \bar{H})$, where $\bar{H}$ denotes the negation of H. Perhaps it is fair to say that Bayes only got half-way to the Bayes factor. This theorem was already familiar to Jeffreys (1939), but without Turing's appealing terminology. The result is especially 'Bayesian' if either H or $\bar{H}$ is composite.

3. Sequential analysis and log factors

Turing was one of the independent inventors of sequential analysis for which he naturally made use of the logarithm of the Bayes factor. He did not know that the logarithm of a Bayes factor occurred in a paper by the famous philosopher Charles Saunders Peirce (1878), who had called it weight of evidence.

To show the relationship to Shannon information it is convenient to write $W(H : E)$ for the 'weight of evidence, or log factor, in favour of H provided by E'. The colon, meaning provided

by, must be distinguished from a vertical stroke, meaning given. As a very slight generalization we naturally define the weight of evidence concerning H as against H', provided by E, by

$$W(H/H' : E) = \log \frac{O(H/H' \mid E)}{O(H/H')} = \log \frac{\text{pr}\,(E \mid H)}{\text{pr}\,(E \mid H')} = W(H : E \mid H \text{ or } H').$$

We then see that weight of evidence is closely related to amount of information concerning H provided by E, $I(H : E)$, defined as $\log\{\text{pr}\,(E \mid H)/\text{pr}\,(E)\}$. In fact

$$W(H/H' : E) = I(H : E) - I(H' : E).$$

The expectation of $I(H : E)$ with respect to H and E, when H and E have a joint probability distribution, is prominent in Shannon's mathematical theory of communication (Shannon, 1948). One can also regard amount of information as a special case of weight of evidence $W(H/H' : E)$ in which H' is replaced by a tautology. In fact weight of evidence is a more intuitive concept than amount of information; and expected weight of evidence, which is an expression of the form $\Sigma p_i \log (p_i/q_i)$, is more fundamental than entropy. It even seems to be advantageous to replace entropy by expected weight of evidence in the proof of Shannon's coding theorems: see Good & Toulmin (1968). Turing's interest in expected weight of evidence will be explained below.

4. The deciban

Turing was the first to recognize the value of naming the units in terms of which weight of evidence is measured. When the base of logarithms was e he called the unit a natural ban, and simply a ban when the base was 10. It was much later that a unit of information for base 2 was called a bit and the same units can be used for information as for weight of evidence. Turing introduced the name deciban in the self-explanatory sense of one-tenth of a ban, by analogy with the decibel. The reason for the name ban was that tens of thousands of sheets were printed in the town of Banbury on which weights of evidence were entered in decibans for carrying out an important classified process called Banburismus.

A deciban or half-deciban is about the smallest change in weight of evidence that is directly perceptible to human intuition. I feel that it is an important aid to human reasoning and will eventually improve the judgements of doctors, lawyers and other citizens.

The main application of the deciban was to sequential analysis, not for quality control but for discriminating between hypotheses, just as in clinical trials or in medical diagnosis.

5. The weighted average of factors

The main application of weights of evidence in 1941 was in situations where H and $\bar{H}$, or H', were simple statistical hypotheses, so that the Bayes factor then reduced to a likelihood ratio. But this was not always so, and sometimes a theorem of 'weighted averages of factors' was relevant (Good, 1950, pp. 68, 71). Turing had noticed a special case of this theorem and the generalization was straightforward.

6. The design of experiments and expected weights of evidence

For evaluating Banburismus in advance Turing calculated the expected weight of evidence. In other words, for this application, he recognized that expected weight of evidence was a criterion for the value of an experimental design. In view of the close relationship between weight of evidence and amount of information, it should be recognized that he partially anticipated unpublished work by L. J. Cronbach, in a College of Education, University of Illinois, Urbana report in 1953, Good (1955–6) and Lindley (1956), all of whom proposed the

use of expected amount of information in the Shannon sense. Of course, Fisher (1925) had used the same philosophical concept much earlier, but with his different definition of amount of information.

Turing remarked that the expected weight of evidence in favour of a true hypothesis is nonnegative, as one would intuitively require. As a mathematical inequality this is a simple result, previously known, for example, to Gibbs (1902, pp. 136–7), but the application to statistical inference is of interest.

7. The variance of weight of evidence

Also while evaluating Banburismus in advance, Turing considered a model in which the weight of evidence W in favour of the true hypothesis H had a normal distribution, say with mean μ and variance σ^2. He found, under this assumption, (i) that if H is false W must again have a normal distribution with mean $-\mu$ and variance σ^2, and (ii) that $\sigma^2 = 2\mu$ when natural bans are used; it follows that σ is about $3\sqrt{\mu}$ when decibans are used. This result was published by Birdsall (1955) in connection with radar, and was generalized by Good (1961) to the case where the distribution of W is only approximately normal. In radar applications the variance is disconcertingly large and the same was true of Banburismus.

8. Expected Bayes factors

Turing noticed a simple and curious property of Bayes factors, namely that the expectation of the Bayes factor against a true hypothesis is equal to unity. This is equivalent to the fundamental identity of sequential analysis (Wald, 1944, p. 285). Wald gave it useful applications that were not anticipated by Turing.

9. Search trees

Closely related to sequential analysis is the concept of a search tree, now familiar in most expositions of decision theory. Such trees occurred centuries ago in games such as chess, and they form part of the technique of cryptanalysis. Certainly Turing made use of search trees, though not with the full explicit apparatus of expected utilities. I do not know whether he had the idea independently of other people or whether it was obvious to many cryptanalysts.

10. The repeat rate

One cryptanalytic idea that I believe Turing had for himself, but which had been anticipated, was that of a repeat rate. If $p_1, \ldots, p_t$ are the mutually exclusive and exhaustive probabilities of the symbols or letters of a t-letter alphabet, occurring in a random sequence, then the probability that two letters in different places will be the same letter of the alphabet is of couse Σp_i^2. Since this is the probability of a 'repeat', Turing called it the repeat rate ρ, an almost self-explanatory term. If, in a sample of N letters, letter i occurs ν_i times, then Turing knew that an unbiased estimate of ρ is $\Sigma\nu_i(\nu_i - 1)/\{N(N-1)\}$. Friedman (1922) had previously in effect called $t\rho$ the index of coincidence. See also Saccho (1951, p. 185). The repeat rate had also been used as a measure of diversity by Gini (1912), according to Bhargava & Uppulari (1975). E. H. Simpson and I both obtained the notion from Turing.

11. Empirical Bayes

Suppose that a random sample is drawn from an infinite population of animals of various species, or from a population of words. Let the sample size be N and let n_r distinct species be each represented exactly r times in the sample, so that $\Sigma r n_r = N$, and n_r can be called 'the frequency of the frequency r'. Turing, using an urn model, showed that the expected

population frequency of a species represented r times is about $(r+1)\,n_{r+1}/(Nn_r)$. For a more exact statement, including the need for smoothing the n_r's, and for numerous elaborations and deductions see Good (1953, 1969) and Good & Toulmin (1956). This work was an example of the empirical Bayes method which method now of course has an extensive literature both with hyperparameterized families of priors and with general priors.

This work was supported in part by a grant from the National Institutes of Health.

References

Bhargava, T. N. & Uppulari, V. R. R. (1975). On an axiomatic derivation of Gini diversity, with applications. *Metron* **33**, 41–53.

Birdsall, T. G. (1955). The theory of signal detectability. In *Information Theory in Psychology*, Ed. H. Quastler, pp. 391–402. Glencoe, Illinois: The Free Press.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700–25.

Friedman, W. F. (1922). *The Index of Coincidence and its Applications in Cryptography*. Geneva, Illinois: Riverbank Laboratories.

Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics*. Reprinted (1960), New York: Dover.

Gini, C. (1912). Variabilità e mutabilità. *Studi Economico-giuridici della Facoltà di Giurisprodenza dell' 'Università' di Cagliari*, III, part II.

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

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–64.

Good, I. J. (1955–6). Some terminology and notation in information theory. IEE Monograph 155R (1955). *Proc. Inst. Elec. Eng.* C **103**, 200–4.

Good, I. J. (1961). Weight of evidence, causality and false-alarm probabilities. In *Information Theory, Fourth London Symposium* (1960), Ed. C. Cherry, pp. 125–36. London: Butterworth.

Good, I. J. (1969). Statistics of language. In *Encyclopaedia of Linguistics, Information and Control*, Ed. A. R. Meetham, pp. 567–81. London: Pergamon.

Good, I. J. & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63.

Good, I. J. & Toulmin, G. H. (1968). Coding theorems and weight of evidence. *J. Inst. Math. & Applic.* **4**, 94–105.

Jeffreys, H. (1939). *Theory of Probability*. Oxford: Clarendon.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.

Peirce, C. S. (1878). The probability of induction. *Pop. Sci. Monthly*. Reprinted (1956) in *The World of Mathematics*, Vol. 2, Ed. J. R. Newman, pp. 1341–54. New York: Simon and Schuster.

Saccho, L. (1951). *Manuel de Cryptographie*. French edn by J. Bres from the Italian. Paris: Payot.

Shannon, E. C. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423, 623–56.

Turing, A. M. (1936–7). On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.* 2, **42**, 230–65; **43**, 544–6.

Turing, S. (1959). *Alan M. Turing*. Cambridge: Heffer.

Wald, A. (1944). On cumulative sums of random variables. *Ann. Math. Statist.* **15**, 283–96.

[*Received October* 1978. *Revised January* 1979]