Casper Albers
Psychometrie & Statistiek
Rijksuniversiteit Groningen
c.j.albers@rug.nl

Column Casper sees a chance

The statistician Alan Turing

Caspers Albers regularly writes a column on everyday statistical topics in this magazine.

This special issue of *Nieuw Archief voor Wiskunde* is devoted to cryptography and it shouldn't therefore be surprising that I devote my column to Alan Turing, probably the most well-known code breaker. Apart from playing a pivotal role in cryptography as well as being one of the founding fathers of computing science, Turing made some important contributions to the theory of Bayesian statistics.

About ten years ago, I was a research fellow at the Open University in Milton Keynes, England. My room was located in a building with the uninspiring name 'M building' and my fellowship was in the department of 'Mathematics, Statistics and Computing'. Very few people made significant contributions to all three of these fields. Even fewer did that in the vicinity of (what is now) Milton Keynes. Only one person ticked both these boxes and saved millions of lives whilst doing so: Alan Turing. After I left the Open University, they renamed the M building after him (I don't think these two events are related).

Turing is well known for his contributions to cryptography during the Second World War, working with other scientists at Bletchley Park and cracking the German Enigma machine. According to Churchill, their work caused the war to end years earlier, thus saving millions of lives. He's one of the handful of mathematicians to have an Oscar-winning movie (*The Imitation Game*) based on his live. His work on the foundations of computing — most notably his papers describing what is now called the Turing Machine [6] and the Turing Test [7] — has been well documented. His contributions to statistics, however, are less well known, in part because this work consisted of classified documents.

During the war, Turing wrote two reports on probability and statistics [8,9], but these remained classified by GCHQ until 2012, seventy years later. Both before and after the release of these documents in 2012, various scientists, including Turing's direct colleagues Edward Simpson (from Simpson's paradox) and Jack Good (known for his work on Bayesian analysis), have reflected on his work in statistics [1,3,5]. Two of the topics he worked on, Bayes factors and sequential analysis, have become standard equipment in the Bayesian statistician's toolbox.

Bayes factors

In null hypothesis significance testing (NHST) the objective is to decide between two competing hypotheses on some parameter of interest. These hypotheses can be simple (e.g. $\mathrm{H}_0\colon \theta = \theta_0$) or composite (e.g. $\mathrm{H}_1\colon \theta \in \Theta_1$, with $\theta_0 \notin \Theta_1$). Jerzy Neyman and Egon Pearson proved in 1933 that the likelihood ratio approach is uniformly most powerful when both hypothesis are simple, $\mathrm{H}_0\colon \theta = \theta_0$ and $\mathrm{H}_1\colon \theta = \theta_1$, but the approach can also be used when the alternative hypothesis is composite. As the name suggests, the likelihood ratio test looks at the ratio of the likelihood under the null distribution and the (maximum) likelihood under the alternative. When this ratio falls below a pre-specified threshold c, the null hypothesis is rejected in favour of the alternative. The likelihood function is given by $L(\theta) = \prod_{i=1}^n f(x_i; \theta)$, and the likelihood ratio is

$$\lambda = \frac{L(\theta_0)}{\max_{\theta \in \Theta_1} L(\theta)}.$$

Example. Let $\underline{x} = x_1, \ldots, x_n$ be a random sample from the $N(\mu, 1)$ distribution. Suppose we are interested in $\mathrm{H}_0\colon \mu = 0$ versus $\mathrm{H}_1\colon \mu \neq 1$. The likelihood function is given by

$$L(\mu \mid \underline{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(x_i-\mu)^2} = (2\pi)^{-\frac{n}{2}} \mathrm{e}^{-\frac{n}{2}(\bar{x}-\mu)^2} \times \mathrm{e}^{-\frac{1}{2}\Sigma_{i=1}^{n}(x_i-\bar{x})^2}.$$

Under H_1, this function is maximised when the maximum likelihood estimator $\hat{\mu} = \bar{x}$ is used. The likelihood ratio can be simplified into

$$\lambda = \frac{L(\theta_0)}{L(\hat{\mu})} = \mathrm{e}^{-\frac{n}{2}(\bar{x}-\mu_0)^2}.$$

Thus, when $n(\bar{x}-\mu_0)$ is larger than a certain value, $\lambda < c$ and the test will reject H_0 in favour of H_1. This frequentist notion of likelihood ratios is turned into a Bayesian one by taking prior information into account. Bayes' Theorem dictates that

$$\underbrace{\frac{\mathrm{P}(\mathrm{H}_0 \mid \underline{x})}{\mathrm{P}(\mathrm{H}_1 \mid \underline{x})}}_{\text{Posterior odds}} = \underbrace{\frac{\mathrm{P}(\underline{x} \mid \mathrm{H}_0)}{\mathrm{P}(\underline{x} \mid \mathrm{H}_1)}}_{\text{Bayes factor}} \times \underbrace{\frac{\mathrm{P}(\mathrm{H}_0)}{\mathrm{P}(\mathrm{H}_1)}}_{\text{Prior odds}}.$$

Thus, the Bayes factor is the ratio of *marginal* likelihoods

$$\mathrm{P}(\underline{x} \mid \mathrm{H}_i) = \int_{\Theta_i} \underbrace{\mathrm{P}(\underline{x} \mid \theta, \mathrm{H}_i)}_{\text{Likelihood}} \times \underbrace{\mathrm{P}(\theta \mid \mathrm{H}_i)}_{\text{Prior}} \,\mathrm{d}\theta \quad (i = 0, 1)$$

Photo credits: Antoine taveneaux, Wikimedia Commons.

Statue of Alan Turing at Bletchley Park

and as such quantifies the evidence in the data for H_0 relative to H_1. When both H_0 and H_1 are simple, the Bayes factor approach is identical to the likelihood ratio approach.

In contrast to the frequentist approach, the Bayesian approach is symmetrical: it can be used to reject H_0 in favour of H_1, but it can also be used to reject H_1 in favour of H_0. Furthermore, the ability to incorporate initial knowledge, or even initial guesstimates, through the prior odds, provide important benefits. Under mild conditions, the Bayesian alternative was easier to compute than the frequentist alternative, an important advantage in the era before modern computing.

Bayes factors thus are a method to quantify the (relative) weight of evidence of two competing hypotheses. Together with Jack Good, Turing introduced the terminology 'deciban' for the units in which weight of evidence was measured. The base-10 logarithm of the Bayes Factor is measured in decibans. Turing and Good's measure of evidence predates the, now more popular, 'bit' by Shannon and Tukey by about eight years. Bayes factors were already known shortly before the war [2], but according to Good [1], Jeffrey's approach lacked the appealing terminology that Turing's approach did have.

Sequential analysis

For Turing, the logarithms of the Bayes Factors were natural ingredients in the sequential analyses. In sequential analysis, the sample size is not fixed in advance: the data are evaluated continuously as they are collected. In a frequentist context, derived by Abraham Wald [10], the usual approach after each new observation is to choose between three alternatives: (1) accept H_0, (2) reject H_0, (3) remain unsure and continue sampling. Deciding between (1), (2) or (3) occurs based on whether a certain likelihood ratio is either below $c_1 < 1$, above $c_2 > 1$ or between c_1 and c_2.

Independently of Wald and his colleagues at Columbia University, Turing derived a Bayesian reasoning for sequential analysis. By taking the prior odds for observation $k+1$ equal to the posterior odds after observation k, and by applying Bayes Factors in each step, he derived a sequential process for updating his degrees of belief. This empirical Bayesian approach is now common within Bayesian statistical research. It has been shown that in many common situations this Bayesian approach coincides with Wald's approach. Even though "Wald gave [sequential analysis] useful applications that were not anticipated by Turing" [1], Turing played an important role in the development of sequential analysis.

He used his sequential conditional probabilities in his work on Banburismus, an automated process to find the most likely settings of the German Enigma machine. This Banburismus was a predecessor of a modern day computing algorithm. In her popular scientific book, McGrayne [4] describes how this system enabled Turing to make a guess of a series of letters in an Enigma message and then to measure his degree of belief in this guess. Subsequently, as additional information came in — e.g. in the form of additionally intercepted Enigma messages — the degree of belief would be updated and the initial guess would be replaced by a guess with a higher degree of belief.

Polymath

With the general public — and probably even within mathematicians — Alan Turing is much better known for his other activities than for his statistical work. However, he deserves a place in a line of polymaths, statisticians that are also well known for contributions to other fields such as Ronald Fisher, Francis Galton and Karl Pearson. The impact of Turing's statistical work has been undervalued, in part due to that many of his contributions were designated as classified information and not released until a couple of years ago. It is the task of the 21st century statistician to correct this omission in the recollection of 20th century statistics.

References

1 I.J. Good, Studies in the History of Probability and Statistics. XXXVII A.M. Turing's statistical work in World War II. *Biometrika*, 66(2) (1979), 393–396.

2 H. Jeffreys, *Theory of Probability*, Clarendon Press, Oxford, 1939.

3 K.V. Mardia and S. Barry Cooper, Alan Turing and enigmatic statistics, in: S. Barry Cooper and A. Hodges, eds., *The Once and Future Turing – Computing the World,* Cambridge University Press, 2016, Chapter 5.

4 S.B. McGrayne, *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy,* Yale University Press, 2012.

5 E. Simpson, A life in statistics – Edward Simpson: Bayes at Bletchley Park, *Significance*, 18 May 2010.

6 A.M. Turing, On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society* 42(2) (1936).

7 A.M. Turing, Computing machinery and intelligence, *Mind* 59 (1950), 433–460.

8 A.M. Turing, Paper on statistics of repetitions (1941), ArXiv:1505.04715v2.

9 A.M. Turing, The applications of probability to cryptography (1942). ArXiv: 1505/04714v2.

10 A. Wald, Sequential Tests of Statistical Hypotheses, *The Annals of Mathematical Statistics* 16(2) (1945), 117–186.