

Editorial

After 20 years of outstanding service, Graham Hoare has retired as Letters Editor for *Mathematics Today*. He brought much expertise and professionalism to this honorary position, attending meetings regularly and responding to letters in a characteristically erudite and jocular style. The Editorial Board is most grateful and wishes him well for the future. Graham celebrated this retirement with a substantial donation to the IMA coffers.

Christopher Hollings (University of Oxford) has kindly agreed to take over the role of Letters Editor forthwith. Elsewhere, Ron Knott (University of Surrey) was recently elected to succeed Edmund Chadwick as North West Branch Chair. Chris and Ron are well known to IMA members for their valued contributions.

Congratulations to Dr Priya Subramanian (University of Leeds), who was awarded a L'Oréal-UNESCO Fellowship for Women in Science this May. At our request, she agreed to write an article for *Mathematics Today*, which you can find on pages 140–141. We also have items about this year's winners of the Royal Society's Copley Medal (Sir Andrew Wiles) and the Shaw Prize in Mathematical Sciences (Professors János Kollár and Claire Voisin).

In October, we shall publish a special issue on the theme of *Space*, as requested in last year's survey, with guest editor Professor Jörg Fliege (University of Southampton). Meanwhile, August's issue includes articles on algorithms, further education, personnel management, centres of gravity and d'Alembert, so there should be something to please everybody.

From September, new AS and A-level syllabi in maths and further maths will be delivered in England, Wales and Northern Ireland. Regrettably, these courses teach significance tests and confidence intervals as the predominant forms of statistical inference, despite their substantial deficiencies. These methods similarly feature in the Scottish advanced higher statistics qualification and in many UK university degree programmes.

The concepts behind significance tests and confidence intervals contributed much to the development of mathematical statistics in the 20th century. However, O'Hagan and Forster [1] note that they have some philosophical flaws, whereas the methodology published by Thomas Bayes in 1763 and Pierre-Simon Laplace in 1814 is fundamentally sound.

Dennis Lindley remarked that significance tests and confidence intervals are statements about data given parameter, rather than parameter given data. They interpret the conditional probability $P(\text{data}|\text{parameter})$ as though it were $P(\text{parameter}|\text{data})$. This misinterpretation can be devastating: the probability of diagnosis given disease is different from that of disease given diagnosis. Similarly, the probability of evidence given innocence is different from that of innocence given evidence. If these probabilities were equal, we should prosecute all lottery winners for fraud!

Other critics of significance tests and confidence intervals include Chair of the Council for the Mathematical Sciences Sir Adrian Smith [2] and President of the Royal Statistical Society



Pierre Simon de Laplace (1749–1827)

Sir David Spiegelhalter [3]. A recent article by Matthews [4] endorses the American Statistical Association's criticisms of p -values for similar reasons, particularly in connection with the current replication crisis. Let us consider the problems in more detail.

When a teacher tests a simple hypothesis about a parameter by collecting data and generating conclusions such as reject H_0 at the 5% level of significance because a 95% confidence interval excludes the test value or because the p -value is less than 0.05, many students are confused by the arbitrary terminology and illogical reasoning.

As probability is the agreed measure of uncertainty, what roles do significance and confidence play? Why choose 5%, 95%

and 0.05 thresholds to make binary decisions? How can some null and alternative hypotheses not partition the set of possible values? Why specify a power of 80% to determine a suitable sample size? Even more perplexing, why perform any statistical analysis when H_0 is often clearly false anyway?

For example, if $H_0 : \pi = 0.5$ where $\pi \in (0, 1)$ is the probability that a coin toss results in a head, the teacher might toss a coin n times and count the number x of heads. Suppose that $n = 10$ and $x = 9$. We estimate $\pi \approx 0.9$ with 95% confidence interval $(0.55, 1.00)$ and p -value about 0.02, so we reject H_0 at the 5% level of significance. However, slight asymmetries in design, manufacture and wear mean that all coins exhibit some bias, so we already know that $\pi \neq 0.5$. The test is ill defined and the conclusion is irrelevant.

What we really wish to determine is whether the coin is approximately unbiased as defined by the parameter $\pi \in A$, where $A = (0.4, 0.6)$ or some other interval of acceptability, given the observed data $D = \{9 \text{ heads}, 1 \text{ tail}\}$. Elementary probability theory does exactly this:

$$P(\pi \in A|D) = \int_A f(\pi|D) d\pi \quad (1)$$

with, from Bayes' theorem,

$$f(\pi|D) = \frac{p(D|\pi)f(\pi)}{p(D)} \propto \mathcal{L}(\pi; D)f(\pi) \quad (2)$$

where $\mathcal{L}(\pi; D) = p(D|\pi)$ as a function of π . For simplicity, probability mass functions p and probability density functions f are distinguishable only by their arguments here.

Relation (2) is cited as 'posterior $\propto$ likelihood $\times$ prior' and is the key to Bayesian inference. The function $\mathcal{L}(\pi; D)$ is the likelihood of parameter π given data D , which is the binomial probability

$$\mathcal{L}(\pi; D) \propto \pi^9(1 - \pi)$$

for our illustration. The function $f(\pi)$ is a prior density that reflects existing knowledge about the coin. My prior beliefs are expressed by the beta density

$$f(\pi) \propto \pi^{25}(1 - \pi)^{25}$$

for $\pi \in (0, 1)$ based on tertiles of 0.47 and 0.53, though you could use a different formula to express your prior beliefs. We need only specify functions that are proportional to the likelihood and prior.

Substituting these two functions of π into relation (2) gives the posterior density

$$f(\pi|D) \propto \pi^{34}(1 - \pi)^{26}$$

for $\pi \in (0, 1)$. This necessarily integrates to one over the unit interval, so the constant of proportionality is

$$\frac{1}{\int_0^1 \pi^{34}(1 - \pi)^{26} d\pi} \approx 4 \times 10^{18}.$$

Finally, equation (1) gives

$$P\{\pi \in (0.4, 0.6)|D\} = \int_{0.4}^{0.6} f(\pi|D) d\pi \approx 0.71$$

and this probability measures my belief that the coin is fair. Unlike a significance test, this analysis involves nothing arbitrary or illogical and generates precisely the answer that we sought!

My prior and posterior densities for π are displayed in Figure 1 and the probability of 0.71 corresponds to the area shaded green. Teachers could use freely available mathematics software to plot graphs and evaluate integrals. Typing `(integral p^34(1-p)^26 from 0.4 to 0.6)/(integral p^34(1-p)^26 from 0 to 1)` into WolframAlpha returns 0.71 almost immediately.

An alternative presentation of this analysis is useful when there is no obvious interval of acceptable values. This involves finding the limits of a 95% posterior probability interval (or union of intervals) B for π and checking whether the test value lies inside B . These intervals are usually defined by

$$\int_B f(\pi|D) d\pi = 0.95$$

and

$$f(\pi_1|D) \geq f(\pi_2|D)$$

for all $\pi_1 \in B$ and $\pi_2 \notin B$. They might resemble confidence intervals but are their antitheses.

For our example, $B \approx (0.44, 0.69)$ includes the test value $\pi = 0.5$, and so supports the hypothesis that the coin is fair.

Evaluating these limits involves numerical solution of nonlinear equations with nested quadrature. Such algorithms are readily available in standard packages, though teachers could use a simpler method. As posterior densities are asymptotically normal, approximate limits are given by $\mu \pm 1.96\sigma$ where $\mu = E(\pi|D) \approx 0.56$ and $\sigma^2 = \text{var}(\pi|D) \approx 0.0039$ here.

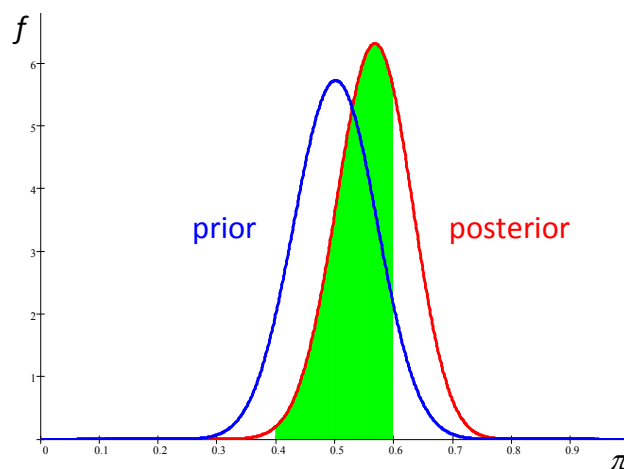


Figure 1: Prior and posterior densities.

All these techniques can be simplified for practical application and have been developed substantially for advanced research. With today's powerful computers, perhaps it is time to stop misleading pupils with significance tests and confidence intervals, and return to the sound methods of Bayes and Laplace.

David F. Percy CMath CSci FIMA
University of Salford

REFERENCES

- 1 O'Hagan, A. and Forster, J. (2004) *Kendall's Advanced Theory of Statistics: Bayesian Inference*, Edward Arnold, UK.
- 2 Bernardo, J.M. and Smith, A.F.M. (2006) *Bayesian Theory*, John Wiley & Sons, Canada.
- 3 Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2012) *The BUGS Book: A Practical Introduction to Bayesian Analysis*, CRC Press, USA.
- 4 Matthews, R.A.J. (2017) The ASA's p-value statement, one year on (with discussion), *Significance*, vol. 14, pp. 38–41.