COMMENTARY

Living in a Bayesian world: scientific deduction through induction

Anurag Agrawal

Every day we wakeup with an existing model of the world in our head, incorporating any new data into the model, with a nip here and a tuck there, permitting us to seamlessly (and subconsciously) estimate likelihood of a variety of events. This system has served us well in understanding the world we live in, but is difficult to apply towards objective, unbiased assessment of scientific problems. Almost by definition, the models we use to understand the world are subjective, emerging inferentially from our cumulative experiences. For example, that the sun rises every day from the east permits us to infer that it should rise every day from the east, and every such observation further strengthens the inference. While this approach, referred to as induction, i.e. where a set of repeated observations allows us to infer or induce a larger relationship, is powerful and natural, it is not suited to making sense of solitary sets of experimental data. Experimental science is supposed to be objective with rigorous testing of hypotheses through well-designed experiments. In this approach, we do not prove the hypothesis to be true, but rather through a series of falsification tests, try to reject the hypothesis. For example, we could set the hypothesis as that the sun may rise from any direction. Within a few days of observation, we could reject this hypothesis since the observations would be extremely unlikely under this hypothesis. Similarly, any hypothesis other than the sun rising in the east, could be rejected. The obvious limitation of this approach is that while it does exceedingly well at identifying anomalies between the data and the hypothesis, it does not necessarily provide a resolution. Blind application of such methods may lead us to reject a hypothesis (typically the null hypothesis) because it appears unlikely without much consideration to the likelihood of the alternate, which unfortunately often remains unstated[1].

Sherlock Holmes is supposed to have said that once you eliminate the impossible, whatever remains, however improbable, must be true. He sees the world in black and white – impossible and possible. Shades of grey corresponding to degrees of probability are ignored. That corresponds to deductive logic where hypotheses are tested and eliminated. An example of a different, more inductive, approach would be of a physician examining a patient. A number of competing hypotheses (diagnoses) emerge at every step of the encounter, starting from watching the patient walk in, to eliciting a medical history, performing the examination and ordering and interpreting laboratory tests. Each piece of information changes the model, dictates the next piece of data required, and all conclusions are nuanced by probabilities. In this approach, assessment of prior probabilities of different diseases is important, corresponding to medical wisdom like ‘common things are common’ and ‘when you hear hoof beats, think horses, not zebras’. Thus a strongly positive syphilis test in a nun could be ignored, even if only 5% of uninfected people have a positive test, if there appears to be no good reason to pursue it further. It would be valid to say that because the disease is extremely unlikely in a nun, it is more likely that the test is a false positive, even though false-positive tests may be unlikely when seen as an independent hypothesis. A different observation in the patient described above, like a tattoo in a private area, may again shift the balance and lead us towards different conclusions, even if the patient claimed to be a nun. Each piece of data acts upon a prior probability distribution to yield a posterior probability distribution. It takes strong new information to significantly shift a strong prior probability. Mathematical structures that capture the essence of this relatively more complex reasoning system exist, generally referred to as Bayesian methods[1–5]. Application of these methods towards scientific deduction is challenging but possible, and will be briefly discussed in the context of common problems in biological sciences.

Thomas Bayes, after whom the Bayesian methods are named, lived more than two and a half centuries ago. For two events A and B, he addressed the relationship between probability of event A given that event B exists [$P(A|B)$], and probability of event B, given that A exists [$P(B|A)$]. This is a common problem in science, where we would like to know the unknown probability of model A, given data B, but instead rely upon calculation of the exact probability of data B, given model A. It should be obvious that the P value, based on which most scientific decisions are made, is a special case of $P(B|A)$, where B is the data distribution and A the null hypothesis. Yet it does not allow us to understand $P(A|B)$, i.e. what is the probability of the null hypothesis being true. To illustrate, $P = 0.05$ only means that assuming null hypothesis to be true, i.e. model A, the probability of a data distribution event similar or more extreme to the one seen, i.e. B, is 5%. To think that this implies that the probability of the null hypothesis being true is 5%, is a common error. The probability of the null hypothesis being true, given any dataset, cannot be calculated by objective Fisherian statistics[2]. It can only be estimated using Bayesian methods that include a subjective component of prior probability distributions[3]. The Bayes theorem states that:

$$P(A|B) \times P(B) = P(B|A) \times P(A),$$

where $P(A)$ and $P(B)$ are the prior probabilities of events A and B.

In the context of an experiment with null hypothesis $H0$, competing alternate hypothesis $H1$, and data D, this can be reformatted as:

$$[P(D|H0)/P(H0|D)] \times P(H0)$$
$$= P(D) = [P(D|H1)/P(H1|D)] \times P(H1),$$
$$[P(H0|D)/P(H1|D)] = [P(H0)/P(H1)] \times [P(D|H0)/P(D|H1)]. \qquad (1)$$

Ratios of probabilities of the null hypothesis versus a competing hypothesis are the same as the odds (or likelihood) of the null hypothesis: prior being before data, and posterior being after data.

Therefore eq. (1) can be restated as – Posterior odds (of null hypothesis) = Prior odds × Bayes factor.

Bayes factor, [$P(D|H0)/P(D|H1)$], simply represents the ratio of the probability of data, similar to that observed, occurring under either hypothesis and permits objective comparison of two hypotheses. Notably, the Bayes factor is independent of any subjective priors, and is possibly superior to the P value in that it forces comparison against alternate hypothesis. In situations where there is no alternate hypothesis, a minimum Bayes factor can be calculated analogous to the P value by assuming an $H1$ which is identical to the data. Since this provides the highest possible probability of $P(D|H1)$, this is the lowest possible value of the Bayes factor, and therefore the strongest possible evidence against the null hypothesis. It can be derived for Gaussian distributions that the minimum Bayes factor is equal to $\exp(-Z^2/2)$. Greater detail of this derivation and the Bayes factor in general can be found in Steven Goodman's review on this topic and is highly recommended[2,3].

At this point, without exhaustive proof it is worth making two points. First, Bayes factors are analogous to P values, and can be used similarly, except that they represent likelihood shifts that do not require correction for multiple comparisons. Second, if prior likelihood of the null hypothesis is high, it will usually remain so except for the most extreme evidence. Each of these is illustrated by examples, especially pertaining to medical genomics.

Examples

A. A study of genetic effects of a particular single nucleotide polymorphism (SNP) finds a SNP in gene X to be associated with asthma. The calculated P value from a chi-square analysis is $P = 0.05$. What can be concluded?

A conventional Fisherian analysis would lead us to reject the null hypothesis and accept that the SNP was associated with increased risk for asthma. If many polymorphisms had been tested, we would have failed to reject the null hypothesis after correction for multiple comparisons, and would have no concluded that no association was found.

A Bayesian analysis would be quite different. We could calculate the minimum Bayes factor as shown above ($Z \sim 1.96$ for $P = 0.05$).

$$\exp(-Z^2/2) = e^{-1.92} = 0.15. \tag{2}$$

From eq. (2), we see that in this case the posterior odds = prior odds × 0.15.

This implies that the unknown prior odds of the null hypothesis being true (prior likelihood $H0$) would be reduced by 85%. That is not very strong, especially considering this is the maximum possible reduction through estimation of a minimum Bayes factor.

Let us see what this means by putting in an empirical value for the prior. Most explorations of new hypotheses are plagued by uncertainty and a 10% probability of there being an effect is an estimate, which can be considered to be quite optimistic in medical genomics. Assuming that there is a 10% probability of there being an effect ($H1$), the prior odds in favour of null hypothesis ($H0$) are $90:10$, i.e. 9. Note that in this case $H0$ and $H1$ are exclusive, with $H1$ being the same as 'not $H0$' or $(1 - H0)$.

Using eq. (2), the posterior odds are $9 \times 0.15 = 1.35$. Since these are odds ($H0/[1 - H0]$), the probability of $H0$ will be $1.35/(1.35 + 1)$, or about 60%. It can be calculated that the prior would have to be $3:1$ or more in favour of $H1$, for the posterior probability of $H0$ to reach 5% or less. There is no requirement of multiple comparison correction in this approach.

B. A Genome Wide Association Study (GWAS) using a 1 million SNP chip, finds a novel SNP to be associated with asthma, with a P value of 10^{-6}. The genomic region has no known function. Another polymorphism in a gene from an asthma-associated pathway has a P value of 10^{-5}. What can we infer?

For extremely low P values, the effect size is considerably stronger and can shift even low prior likelihoods to a believable zone. In this example, a P value of 10^{-6} or less, typically used for GWAS, corresponds to a z of 4.75 and a minimum Bayes factor of approximately 10^{-5}. Thus prior odds of as low as $1:9999$ (1 in 10,000) towards an effect (i.e. $9999:1$ in favour of the null hypothesis) would yield posterior odds of about $8:1$ in favour of the effect. Yet this is by no means conclusive when the priors are that low, since probability of null hypothesis being true would still be around 11%.

In contrast, if we started with prior odds of $1:99$, even the weaker P value of 10^{-5} ($z = 4.25$) would translate to posterior odds of $99:1$, i.e. probability of the null hypothesis being true would be 1% or less. Clearly, the second situation is superior to the first, for rejecting the null hypothesis, despite the weaker P value.

In modern GWAS, 1 million SNPs may be tested at one time, with priors for some genes being much lower than those described above. Even P values of 10^{-7} are insufficient for firm conclusions. Therefore, today replication in an independent cohort is considered a minimum standard. From a Bayesian viewpoint, the posterior probability of the first study can provide the justification for a higher prior for the second study and deductions can be made with more confidence.

While a detailed description of the appropriate approaches to calculate priors is beyond the scope of this note, many high-quality reviews exist on the subject of Bayesian statistics for GWAS[4,5]. In the previous example, the gene with functional evidence of involvement in asthma but a weaker P value was more likely to be truly associated with asthma risk, because of a stronger prior probability. However, the novel association from an unexplored genomic region may be much more valuable in yielding new insights. Both are important discoveries, as long as they are understood in the proper context.

In summary, Bayesian approaches provide an important perspective for understanding scientific data. Sometimes the directions in which they point can be surprising, especially if we have become used to letting isolated statistical tests guide our understanding. Whether this is good, bad, or unnecessary, the question is open for debate!

1. Howson, C. and Urbach, P., *Scientific Reasoning: The Bayesian Approach*, Open Court, La Salle IL, 1989.
2. Goodman, S. N., *Ann. Intern. Med*., 1999, **130**(12), 995–1004.
3. Goodman, S. N., *Ann. Intern. Med*., 1999, **130**(12), 1005–1013.
4. Knight, J., Barnes, M. R., Breen, G. and Weale, M. E., *PLoS One*, 2011, **6**(4), e14808.
5. Huang, H., Chanda, P., Alonso, A., Bader, J. S. and Arking, D. E., *PLoS Genet*., 2011, **7**(7), e1002177.

Anurag Agrawal is in the Centre of Excellence for Translational Research in Asthma and Lung Disease, CSIR Institute of Genomics and Integrative Biology, Mall Road, Delhi University Campus, Delhi 110 007, India.
e-mail: a.agrawal@igib.in