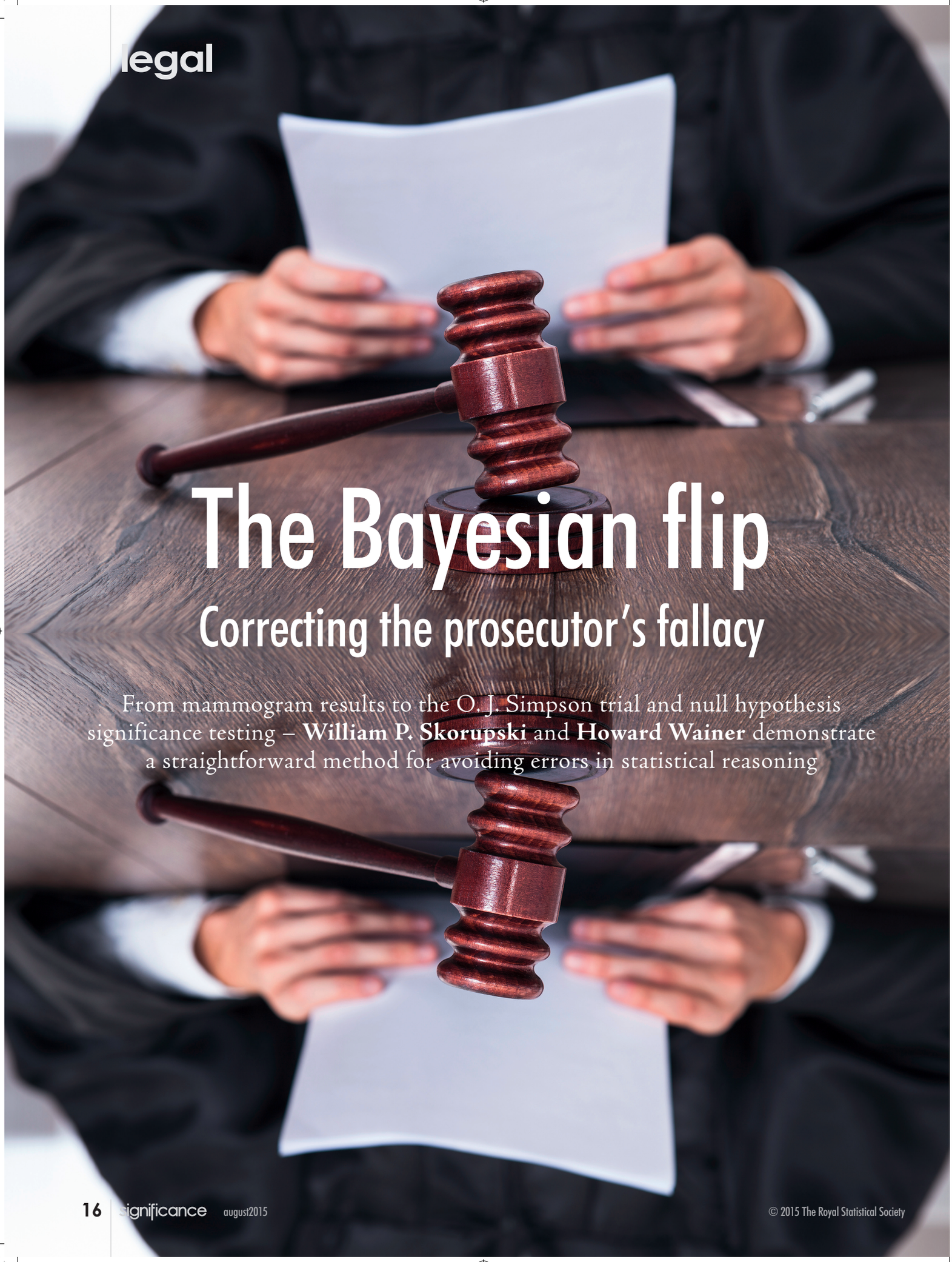


The Bayesian flip

Correcting the prosecutor's fallacy

From mammogram results to the O. J. Simpson trial and null hypothesis significance testing – **William P. Skorupski** and **Howard Wainer** demonstrate a straightforward method for avoiding errors in statistical reasoning

It is rare that a case before the US Supreme Court hinges on a subtle statistical issue, yet that is precisely what occurred on 11 January 2010 in the case of Troy Brown – a man convicted of the rape of a 9-year-old girl. The evidence of Brown's guilt, excluding DNA, was both circumstantial and equivocal. However, the jury's guilty verdict was influenced at least in part by the prosecution's claim that only one in 3 million random people would have the same DNA profile as the rapist, and hence there was only a 0.000 033% chance that Brown was innocent.

Upon appeal of the case, the defence argued that the conclusions drawn from the statistics cited by the prosecution were incorrect, and were an example of what Thompson and Shumann[1] famously called "the prosecutor's fallacy". The Supreme Court, writing in its decision on the Brown case, described the fallacy as:

> the assumption that the random match probability is the same as the probability that the defendant was not the source of the DNA sample. … ("Let P equal the probability of a match, given the evidence genotype. The fallacy is to say that P is also the probability that the DNA at the crime scene came from someone other than the defendant"). … It is further error to equate source probability with probability of guilt. (bit.ly/1BcLATk)

Restating this both more succinctly, and in terms better suited to a statistically literate readership, the prosecutor's fallacy is to calculate P(evidence | innocence) and interpret it as P(innocence | evidence). It may be true that if the accused were innocent, there is only one chance in 3 million of a DNA match. But the DNA match does not necessarily imply that there is only one chance in 3 million of the accused being innocent. Stated more generally, the prosecutor's fallacy is

$$P(A \mid B) = P(B \mid A) \quad (1)$$

We know, from Bayes' rule, that

$$P(A \mid B) = P(B \mid A)\, P(A)/P(B) \quad (2)$$

Equation (1) is only true when the marginal probabilities are equal, $P(A) = P(B)$, or in the Brown case, when the unconditional probability of the observed data is equal to the unconditional probability of innocence. Such situations are not commonly found in practice.

In many scientific and decision-making investigations we may be presented with P(data | theory), the probability of observing what we have observed conditional on an unknown – the truth of our theory. Our theory may be the innocence or guilt of a defendant, or perhaps the status of a null versus alternative hypothesis test. But what we would really like to know is the probability that our theory is correct, given what we have observed: P(theory | data). But how are we to compute this? From the earliest statistics course, in which we assess the fairness of flipped coins, we calculate the probability from the theory – if the coin is fair P(heads) = P(tails) = 0.5. We ascertain, under this model of fairness, how likely it is that we would get the observed proportion of heads to tails, P(heads | fairness). If we confuse that with P(fairness | heads), we are guilty of this statistical fallacy.

In many investigations we may be presented with P(data | theory), but what we would really like to know is P(theory | data): the probability that our theory is correct, given what we have observed

Even a cursory glance through the social science literature shows that prosecutors are not unique in drawing incorrect inferences from probabilities provided by traditional null hypothesis (H_0) tests. In the 1920s, R. A. Fisher laid out the basic ideas of the approach that has come to be called the "likelihood principle".[2] Assuming a sampling distribution for the data, given an underlying theory, is frequently treated as a way to test the plausibility of the theory. However, this kind of thinking can get us into trouble if we are not careful. It is typical for scientific papers to report findings that are "statistically significant" in the results, and then treat those results as evidence of the alternative hypothesis in the discussion. Few say anything as precise as "these data would be unlikely if the null hypothesis were true". Too often we see P(data | H_0) reported and treated as if it were P(H_0 | data). Why the switch?

The well-known Bayesian statistician, Melvin Novick famously referred to P(data | H_0) as a "poor answer to an uninteresting question". Novick's observation is echoed by usage. Indeed, in February this year the journal *Basic and Applied Social Psychology* went so far as to ban the use of null hypothesis testing in favour of descriptive measures such as effect sizes. While the ban does permit the use of Bayesian inference on a case-by-case basis, we are prepared to make a stronger argument in favour of these methods. The problem that persists is that users of statistics would like P(H_0 | data) but do not know how to get it. In some very real sense they have heeded the advice given by Stephen Stills in his 1970 song: "If you can't be with the one you love, love the one you're with." Perhaps getting the right answer was too difficult in the distant past, but not today. There is no longer any reason to settle for the P you're with; you can be with the P you love.

Introducing the Bayesian Flip

To move from P(data | theory) to P(theory | data), we need to do the Bayesian flip. We illustrate the importance of this with a vital example.

Every year in the United States 38 million women are tested for breast cancer with mammograms. Of these, 140 000 have cancer. Mammograms have been determined to be 90% accurate for women with breast cancer. This figure was calculated by tallying all of the women who were eventually determined to have breast cancer and looking back to see if their initial mammograms were positive, thus: P(+mammogram | cancer) = 0.90 and, using a similar empirical investigation, P(+mammogram | no cancer) = 0.10. These two values are referred to as sensitivity (or power) and the false positive (or Type I error) rate, respectively. That they add up to 1.00 is a coincidence; they are not, in general, complementary. It is important to know that a test is both powerful and has a relatively low rate of false positives. But when one is faced with a positive mammogram result, these are hardly useful. We administer a mammogram because we do not know whether or not someone has cancer. What we want to know is

$$P(\text{cancer} \mid +\text{mammogram}) \quad (3)$$

Image: AndreyPopov/iStock/Thinkstock

We can calculate what we want from what we know using equation (2), but for now we will take a short-cut and recognise that the probability in (3) is a fraction that has as its numerator the number of women annually diagnosed with breast cancer via mammograms, or 140 000, and as its denominator the number of positive mammograms (including both true cancer cases and false positives):

$$\mathrm{P}(\text{cancer} \mid +\text{mammogram}) = \frac{\text{True positives}}{(\text{True positives} + \text{False positives})}$$

$$= 140\,000 / (140\,000 + 0.1 \times 38 \text{ million})$$
$$= 140\,000 / (140\,000 + 3\,800\,000)$$
$$= 140\,000 / 3\,940\,000 = 0.036 = 3.6\%$$

Thus, if an asymptomatic woman receives the dreadful news that her mammogram has come back positive, more than 96% of the time it is a false positive – she is fine. The dramatic difference between the 90% statistical power of the test and its 3.6% accuracy demonstrates the importance of not confusing the former with the latter; we must do the Bayesian flip.

The defence attorney's fallacy: the O. J. Simpson trial

Thompson and Shumann[1] named both the "prosecutor's fallacy" and the "defence attorney's fallacy"; both sides of the aisle have engaged in this erroneous thinking. A famous court case from recent history that used probability to discuss the value of evidence was the trial of O. J. Simpson.

In 1994, O. J. Simpson, an actor and former professional American football player, was accused of murdering his ex-wife, Nicole Brown Simpson. News of this court case filled the media for months, so we can skip the details. Of relevance here is that during the trial, Alan Dershowitz, an advisor to Simpson's defence attorneys, claimed that Simpson's previous accusation of spousal abuse was not particularly relevant. The evidence was that only about one in 2500 men who batter their significant others (wives, girlfriends) go on to kill them.

Dershowitz's statement, though backed by evidence, was probability misapplied. In a 1995 letter to the editor in the magazine *Nature*, the prolific statistician I. J. Good[3] pointed out the more relevant concern: if a previously battered woman has been murdered, what is the probability that her batterer committed the crime? Good, using a few simplifying assumptions, expressed the chances of this as roughly one in 3. Clearly, the previous accusations of battery would be considered relevant evidence if we knew that one in 3 murdered women were murdered by their batterers.

Fuse/Thinkstock

Let us redo this argument without simplifying assumptions, but instead by using 1992 data and calculating the relevant probability directly using Bayes' rule. The data for this exercise were taken from the Clark County Prosecuting Attorney webpage on domestic violence (bit.ly/1u32oIq) and a 2010 *New York Times* article by Steven Strogatz.[4] All calculations are using data for women only.

In this example, the notation B represents "woman battered by her husband, boyfriend, or lover", M represents the event "woman murdered", and by extension, M, B denotes "woman murdered by her batterer". Our goal is to compare $\mathrm{P}(M, B \mid M)$ to $\mathrm{P}(M, B \mid B)$.

In 1992, the population of women in the United States was approximately 125 million. That year, 4936 women were murdered. So, one marginal probability, $\mathrm{P}(M) = 4936/125\,000\,000 = 0.000\,04$, or about one in 25 000. Approximately 3.5 million women are battered every year, so we estimate $\mathrm{P}(B) = 0.028$ (3.5 million divided by 125 million). That same year 1432 women were murdered by their previous batterers, so the marginal probability of that event is $\mathrm{P}(M, B) = 1432/125\,000\,000 = 0.000\,01$, or one in 87 290, and the conditional probability, $\mathrm{P}(M, B \mid B)$ is 1432 divided by 3.5 million, or 1 in 2444. These are the numbers Dershowitz used to obtain his estimate that about 1 in 2500 battered women go on to be murdered by their batterers.

Our goal is to calculate the probability of a murdered woman being murdered by her batterer. Using Bayes' rule, we have

$$\mathrm{P}(M, B \mid M) = \mathrm{P}(M \mid M, B)\,\mathrm{P}(M, B) / \mathrm{P}(M)$$

The conditional probability in the numerator, $\mathrm{P}(M \mid M, B)$, is simple enough, for the probability of a woman being murdered, given she has been murdered by her batterer, is 1. So, $\mathrm{P}(M, B \mid M)$ is just the ratio of two probabilities: $\mathrm{P}(M, B \mid M) = \mathrm{P}(M, B)/\mathrm{P}(M) = 0.000\,01/0.000\,04 = 0.29$, or about 1 in 3.5 – slightly lower than I. J. Good's estimate, but obtained using hard data without the need for his simplifying assumptions.

Alan Dershowitz provided the jury with an accurate but irrelevant probability. A murdered woman having been murdered by her batterer is 709 times more likely than a battered woman being murdered by her batterer: $\mathrm{P}(M, B \mid M) \approx 709 \times \mathrm{P}(M, B \mid B)$. The reasoning is subtle and so we must take care in navigating such waters, but the reward is great.

Of course, we could take the same pathway that we followed with the mammogram example, and estimate the probability directly. We note that in 1992 there were 4936 women murdered, of whom 1432 were murdered by men who had previously battered them. Thus $P(M, B \mid M) = 1432/4936 = 0.29$. We went through the Bayesian flip to make the mechanics explicit. The component pieces will not always be so readily obtained.

Brown revisited

Let us return, finally, to the case of Troy Brown, our motivating example, and consider what the inferences would have been had the prosecution used Bayes' rule. We use the notation "DNA" to represent the event "DNA evidence matches the accused", I to represent "innocence", and G to represent "guilt".

Our goal is to compare $P(I \mid \text{DNA})$ to $P(\text{DNA} \mid I)$. The data for this example were obtained from the Oyez Project website (bit.ly/1fbDnEP), an online archive for the Supreme Court.

Recall that the prosecution claimed the random match probability for the DNA evidence was one in 3 million, or 0.000 000 33. That number represents the probability of the DNA found at the crime scene randomly matching that of Brown if the DNA did not in fact come from him (i.e., $P(\text{DNA} \mid I)$). That is one of the three pieces we need to estimate $P(I \mid \text{DNA})$.

The other term in the numerator is $P(I)$, the marginal probability that Brown did not commit the crime. In this context we may consider this probability as the "weight of other available evidence". If there were only two possible perpetrators, each equally culpable in prospect, innocence and guilt would be equally likely and we could choose 0.5 for this probability, thereby making $P(I) = P(G)$.

If we consider the prosecution's case to be largely circumstantial (Brown lived in the same trailer park as the victim, but did not completely match the victim's description of her attacker), we could choose a larger number for $P(I)$, say 0.9. That would represent only a one in 10 chance of guilt without considering DNA evidence.

Or, if there were no other evidence besides the DNA, the presumption of innocence would suggest that the 1 million post-pubertal males within a day's travel would be equally likely suspects. In this case $P(G)$ would be one in 1 million, or 0.000 001.

In this example, we will consider all three probabilities.

The last piece is the denominator, the marginal probability of the DNA evidence, P(DNA). This represents the probability of finding such a DNA match, given the innocence or guilt of the accused (i.e., it represents all the ways we might have obtained this evidence; it is either there because Troy Brown is guilty, or because he is innocent and the DNA randomly matches his DNA):

$$P(\text{DNA}) = P(\text{DNA} \mid I)P(I) + P(\text{DNA} \mid G)P(G)$$

$P(\text{DNA} \mid I)$ has already been estimated by the genetics experts as 0.000 000 33 (we will return to the reasonableness of this estimate shortly). $P(I)$ is a value we have chosen to represent the "weight of other evidence" (we will use 0.5, 0.9 and 0.000 000 1 as examples) and $P(G)$ is its complement, $1 - P(I)$.

To solve for P(DNA), we only need to estimate $P(\text{DNA} \mid G)$. This one is easy. If the accused is guilty, by definition the probability of his DNA matching the source is 1! So, if we use 0.5 for the probability of innocence and guilt, we get $P(\text{DNA}) = 0.000\,000\,33 \times 0.5 + 1 \times 0.5 = 0.500\,000\,167$. To all intents and purposes, if innocence and guilt are

Table 1. Calculation details of P(DNA)

P(DNA)	=	$P(I)$	×	$P(\text{DNA} \mid I)$	+	$P(G)$ ×	×	$P(\text{DNA} \mid G)$
0.5	=	0.5	×	0.000 000 33	+	0.5	×	1
0.1	=	0.9	×	0.000 000 33	+	0.1	×	1
0.000 001	=	0.999 999	×	0.000 000 33	+	0.000 001	×	1

Table 2. Calculation details of $P(I \mid \text{DNA})$

$P(I \mid \text{DNA})$	=	$P(\text{DNA} \mid I)$	×	$P(I)$ /	/	P(DNA)
0.000 000 3	=	0.000 000 33	×	0.5	/	0.500 000
0.000 003 0	=	0.000 000 33	×	0.9	/	0.100 000
0.25	=	0.000 000 33	×	0.999 999	/	0.000 001

equally likely outcomes, then P(DNA | I) is small enough to be ignorable, and so P(DNA) is 0.5. This represents the rare case where the two marginal probabilities are equal (or close enough to equal), P(I) ≈ P(DNA). In this case, the prosecutor gets lucky, because P(DNA | I) = P(I | DNA) = 0.000 000 33.

However, as previously stated, innocence and guilt do not appear equally likely in this case, as all other available evidence was circumstantial. In that case, if we assume that the marginal probability of innocence is much higher, P(I) = 0.9, we get a fairly different result. If we substitute 0.9 for P(I) and therefore 0.1 for P(G), we get P(DNA) = 0.1 (again, P(DNA | I) is small enough to be ignorable). Now that P(DNA) does not equal P(I), the result we obtain for P(I | DNA) is 0.000 003, or one in 333 333. That is still a small chance of being innocent, given the evidence, but it is a number that is nine times larger than the value claimed by the prosecution.

Finally, if there were no other evidence than the DNA match and the presumption of innocence means that there were a million other males close enough to be the perpetrator, the probability of Brown's innocence, given the DNA is 0.25. And yes, that still means that he is three times more likely to be guilty than innocent, but 25% surely implies a reasonable doubt.

A summary of the calculations for P(DNA) is given in Table 1 and the parallel calculations for P(I | DNA) are in Table 2.

Obviously, the choice of prior for P(I) is seriously consequential; any accumulated evidence that affects our estimate of P(I) matters. For example, over the course of the trial, it was revealed that two of Brown's brothers lived in the same trailer park as the victim, and Brown had two other brothers who lived within a 500-mile radius of the crime scene.

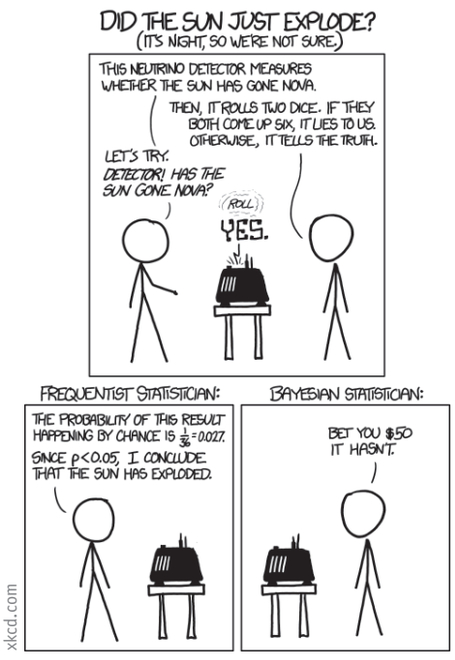
It does not take a genetics expert to realise that two brothers are much more likely to have similar DNA than two randomly sampled individuals from the population. The probability of any two brothers having matching DNA depends on a number of unobserved genetic factors, but it could be as high as 1 in 66, P(DNA | I) = 0.015. This number is still fairly small, but considerably larger than one in 3 million. If we use 0.015 for P(DNA | I) and continue to use P(I) = 0.9, we estimate the P(I | DNA) to be 0.12, certainly enough for a reasonable doubt.

Priors aren't data, or are they?

Bayes' rule is a powerful tool for calculating the P you want to know. It includes a mechanism for weighting the strength of evidence in the likelihood part of a formula. When calculating P(theory | data) and having been given P(data | theory) – the likelihood of the data given the theory – we know that we must learn or estimate the prior P(theory) to do the Bayesian flip.

In the mammogram and O. J. Simpson examples, all of the components could be estimated from data, but we are not always so fortunate. In the Brown case we had to estimate the "weight of other evidence" to complete our calculations. What happens when we try to make this inference without having any idea about the incidence rate?

In most decision-theoretic procedures, the possible events in P(theory) are mutually exclusive and exhaustive. The job of the Bayesian inference-maker is to take the information provided by the data in P(data | theory) and temper it with what is known or not known about P(theory). In the cartoon below (xkcd.com/1132), the Bayesian statistician will wager $50 that the sun has not gone nova because there is *considerable* evidence beyond the dice roll to suggest it has not. In other examples, the two complementary probabilities in P(theory) may be equal, indicating no preconceived notions about which outcome is more likely. We do not have to proceed with only one estimate of P(theory); a range of probabilities may be supplied, providing us with a natural way to create a range of plausible posterior probabilities for P(theory | data).



When considering the weight of other evidence (without DNA) in the Brown case, we recognised that innocence was more likely than guilt, and thus we could weight the value of the DNA evidence accordingly. In the O. J. Simpson example, the prior probability of a woman being murdered by her batterer was directly estimable from other data sources, and we observed that multiple pathways led us to the same conclusion about how likely it is that a murdered woman had been killed by her batterer.

The unifying theme is that the prior P(theory) is used to refine our posterior inferences about P(theory | data). It is not to be lamented or ignored. And, as more information is made available to us, we can use previous posterior probabilities as new priors in our estimation. We can correct the fallacy, whether made by a prosecutor, defence attorney, or social scientist, and be with the P we love.

Acknowledgement

We are especially grateful to Eric Bradlow for his valuable comments on an earlier draft of this paper.

References

1. Thompson, E. L. and Shumann, E. L. (1987) Interpretation of statistical evidence in criminal trials. *Law and Human Behavior*, **11**(3), 167–187.
2. Birnbaum, A. (1962) On the foundations of statistical inference. *Journal of the American Statistical Association*, **57**(298), 269–326.
3. Good, I. J. (1995) When batterer turns murderer. *Nature*, **375**, 541.
4. Strogatz, S. (2010) Chances are. *New York Times*, 25 April. nyti.ms/1HIs2ao

William P. Skorupski is associate professor and programme coordinator for research, evaluation, measurement and statistics at the University of Kansas

Howard Wainer is currently distinguished research scientist at the National Board of Medical Examiners. His most recent book (his twentieth) is *Medical Illuminations: Using Evidence, Visualization & Statistical Thinking to Improve Healthcare* (Oxford University Press, 2014)