Invalid inversion

Stephen Senn warns us that some things work backwards, but others do not.

"You might just as well say," added the March Hare, "that 'I like what I get' is the same thing as 'I get what I like'!"
(Lewis Carroll, *Alice in Wonderland*)

Downside up?

Do we learn from making mistakes or do we make mistakes from what we learn? Well, no doubt we do both because in the words of Artemus Ward, "It ain't so much the things we don't know that get us in trouble. It's the things we know that ain't so." So certainly some of what we have learned we have learned in error and this will be the source of mistakes, but this is a negative and undesirable feature of learning and surely not the same as the desirable improvement attendant upon observing mistakes, searching out their cause and eliminating them from future conduct. So it is all very elementary and all very obvious that mistakes made leading to increased knowledge is not the same as increased knowledge leading to more mistakes. You could scarcely function as a speaker of English, or indeed any other language, if you thought that you could switch subject and predicate around without radically changing your meaning.

However, rather surprisingly perhaps, when it comes to probability statements many people are habitually careless when it comes to reversing the order of things. In so doing they commit a fallacy that is sometimes given a rather grand name; it is called the *error of the transposed conditional*[1]. I prefer to call it *invalid inversion* and I consider that of all the various things one could learn to recognise about probability statements about the real world, this is the most valuable.

Careful with your conditionals

Most women are not breast cancer sufferers, thank goodness, but one would be foolish to conclude that therefore most breast cancer sufferers are not women. This is a particular example I used to use to try and make the fallacy instantly comprehensible to the medical students I used to teach. An example like this was much easier for them to grasp than the abstract formulation of a particular important principle I was trying to impart, namely that the probability of the hypothesis given the evidence is not the same as the probability of the evidence given the hypothesis.

The probability that a woman has breast cancer is small. The probability that someone with breast cancer is a woman is large. Obvious – but some invalid inversions are not so easily seen

Actually, I find that to explain the difference between conditional probabilities it helps to consider some data, and for the United Kingdom these might look something like Table 1, which gives numbers of individuals (in thousands) cross-classified by sex and disease status. These are based on some very rough calculations and should not be taken too seriously, but nevertheless will do for the purpose of making the logical point. It has been estimated that breast cancer prevalence among women in the UK is about 550 000, which is to say that there are this many women alive in the UK who have had breast cancer at some time. It has also been estimated that for every 1000 incident cases of breast cancer among women there are about 6 amongst men. If we make the not necessarily reasonable assumption that relative prevalence reflects relative incidence (since survival rates could be rather different), we can combine it with estimates for the total UK population by sex to get figures such as in Table 1.

Now, if we randomly select an individual from the UK population, the chance that this person belongs in the upper left-hand corner of the table, that is to say, is a female who has or has had breast cancer, is $550/61\,792 \approx 0.008\,90$, which is fairly small. This is an example of what is called a *joint probability* since two conditions (female and breast cancer) are jointly involved. The probability of a randomly chosen individual being a breast cancer victim, however, is an example of a *marginal probability*. It is obtained by dividing the marginal total of breast cancer sufferers (male or female) by the total number of persons. Thus we have $553/61\,792 \approx 0.008\,95$, similarly small, and only a tiny bit different from our previous answer because only a very small number of men have been added into the calculation.

Table 1. Possible figures (in thousands) for breast cancer prevalence in the UK (based on figures by Cancer UK available at `http://info.cancerresearchuk.org/cancerstats/types/breast/incidence/`)

	Sex		
Health status	*Female*	*Male*	*Total*
Suffering from breast cancer	550	3	553
Not suffering from breast cancer	30 868	30 371	61 239
Total	31 418	30 374	61 792

On the other hand, the probability of a given breast cancer sufferer being female is an example of a *conditional probability*. It is obtained by conditioning on a person being a breast cancer victim. That is to say, the table is reduced to the first row only. The total of the two entries in the row (in thousands) is 553, so that now the relevant probability is $550/553 \approx 0.995$, which is very large. You could fairly safely bet on it.

Margin of error

Thus, as we all instantly recognise, the probability of a breast cancer victim being female is extremely high. How could one confuse this with the probability of a female having breast cancer? To calculate the latter we divide the number of female breast cancer sufferers by the number of females. The same figure of 550 used previously appears in the numerator, but now the denominator is not the total of row 1 but the total of column 1. We thus calculate $550/31418 \approx 0.0175$, a much lower figure. Thus it is the denominator, the number below the line, that differs in the two cases, being (in thousands) 553 in the one case and 31 418 in the other.

Thus the probability of a breast cancer victim being female is 550/553; whereas the probability that a female has breast cancer is 550/31 418. It is the fact that the number of females is so much greater than the number of breast cancer sufferers that makes the latter probability small even though the former is large.

> The probability of an abused woman being murdered is low. But if she is murdered, the probability that it was her abuser who murdered her is high. O. J. Simpson's prosecutors missed that

A fist full of fallacies

This is all very well, the reader may say, but does it happen in practice? Is it the case that mistakes of this sort are commonly made? Is invalid inversion a common error? Unfortunately, the answer is "yes". I now consider some famous, in fact notorious, cases.

The first of these concerns the trial of O. J. Simpson for the murder of his wife Nicole Simpson and her friend Ronald Goldman in 1994. O. J. Simpson had been convicted of abuse of his wife in 1989. At the trial for murder, one of the lawyers for the defence, Alan Dershowitz, pointed out to the court that the probability of a woman who had been in an abusive relationship being murdered by her abuser was low. In fact, he estimated the risk to be 1 in 2500. This may well be so, but is it relevant? After all, most women are not murdered, even those in an abusive relationship; this is why the probability is so low. However, since Nicole Simpson *was* murdered, the probability is surely not directly relevant. After all, the probability of a given person winning the lottery is extremely small but if they have the winning ticket, then they have won notwithstanding the low probability. So a more relevant question would seem to be, if a murdered woman has been in an abusive relationship, what is the probability that her murderer abused her? A number of statisticians tried to answer this, including the famous Bayesian Jack Good, who worked with Alan Turing at Bletchley Park during the Second World War. Good initially calculated that this might be as high as 50%[2], and later revised this to 80%[3] when he realized that Dershowitz's initial probability was an annual and not a lifetime risk.

The second concerns the trial of solicitor Sally Clark for murdering two children who had died while young infants. She was convicted in 1999 partly on the grounds that the probability of two cot deaths was extremely small: a figure of one in 73 million was calculated. Actually the probability was not calculated correctly since a one in 8500 chance had been squared to produce it, a calculation that would only be reasonable if independence applied. Note, however, that if one were to apply the logic of the O. J. Simpson case and if one were to accept the one in 8500 probability, one could argue that the first death was irrelevant to judging the second since there is only a very small chance that, one death having happened, it would be followed by another!

The point is, however, that whatever the true probability (which we may suppose to lie somewhere between 1 in 8500 and 1 in 73 million), it is (at best) a probability of the evidence given innocence and not the probability of innocence given the evidence. (This particular error now has a name in the legal literature on evidence. Thanks to a paper by Thompson and Schumann in 1987 it is known as the *prosecutor's fallacy*[4].) Alas, the defence counsel appear to have missed a trick as they could have claimed that there is only a 1 in 10 000 chance (say) of a mother killing one child and hence only a 1 in 100 million chance of her killing two. Sally Clark was sent to prison and only released three years later on appeal.

Slippery simulation

Are statisticians immune from invalid inversion? Since I am a statistician I wish I could claim that we are, but statisticians are also caught out from time to time and I daresay that I am no exception. One area in which blunders have been made is in simulating results. It is very common to examine certain inferential procedures by setting certain values into a model, simulating some data from them and then seeing whether the statistic actually observed is on average equal to the parameters produced by the simulation. If it is, the statistic is said to be *unbiased*.

This is, however, a weasel word. The fact that a statistic $\hat{\theta}$ is on average equal to the parameter θ, the "true" value of which it is an estimate, does not mean that the parameter is on average equal to the statistic. If A is equal to B, then B must be equal to A; but if A is *equal on average* to B then it is *not* true that B must be *equal on average* to A. This is the crux of the matter. The mathematical statement "is equal to" is reversible; the statement "is equal on average to" is not. In mathematical notation the former involves a statement of the form $e[\hat{\theta} \,|\, \theta] = \theta$ – the expectation of the statistic given the parameter is equal to the parameter; the latter is a statement of the form $e[\theta \,|\, \hat{\theta}] = \hat{\theta}$ – the expectation of the parameter given the statistic is equal to the parameter; and in fact the former does not guarantee the latter. Does this matter? Isn't the former what is really more important? Not necessarily. Note that our definition of an unbiased statistic above includes that non-reversible phrase "is on average equal to"; and statisticians have too often reversed it. There have been occasions on which statisticians have used simulation to investigate whether a particular approach to inference is satisfactory or not and have failed to note that being satisfactory in the forward sense is not equivalent to being satisfactory in the backward sense, but the backward sense may be what matters.

> If *A* is "equal on average" to *B*, it does *not* follow that *B* is 'equal on average' to *A*

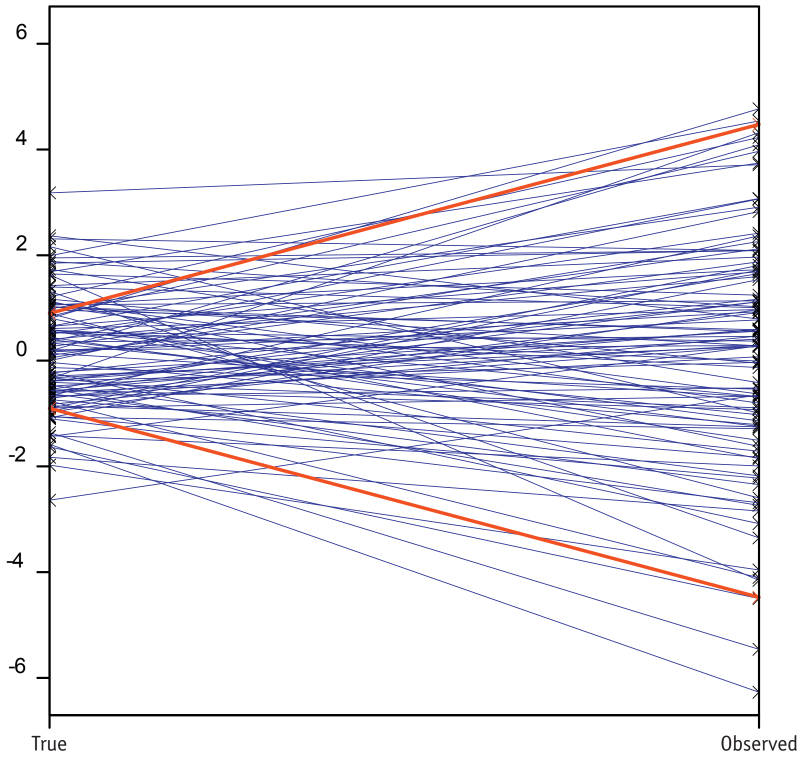

Figure 1. Illustration of the result of a simulation. "True" values are generated from a distribution with mean 0 and variance 1. These are plotted on the left-hand vertical axis. Random noise with a variance of 4 is added and the resulting "observed" values plotted on the right-hand axis. Blue lines join corresponding true and observed values. Two red lines are added to show theoretical *shrinkage* for two observed values. Note the average true value is about ¾ unit away from 0; the average observed value is about 1.5 units away from 0. The average departure point in Figure 1 is at 0. But the departure point, on average, is not at zero distance from 0

Suppose, for example. that you wished to devise an approach to analysing treatment effects from a randomised clinical trial that was conditionally unbiased for a given difference of some baseline characteristic; that is to say, averaged over all trials that had the given baseline, the observed difference would produce the "correct" answer. Due to measurement error the observed difference would not be the true difference. A method that was unbiased for observed differences would not be unbiased for true differences, and one could show this by simulation. However, this would simply be the wrong thing to show. A statistician is never faced with a true difference to deal with but only ever an observed difference. His task is to provide a reasonable answer given what has been observed. The point that is often overlooked here is that although an observed difference will be equal on average to the true difference (if measurement error is random) the reverse is not the case. The true difference will not be equal on average to the observed difference. On average true differences will be closer to the overall mean.

To see this, consider Figure 1. Here I have simulated some true values and added a great deal of random noise to produce some observed values. The true values are spread about the left-hand vertical axis and the observed values on the right-hand side. I have joined the pairs of values using blue lines. You can regard the trajectories as being those produced by a very inexpert billiards, pool or snooker player who has the ball spotted somewhere along the left cushion and is given the task of sending the ball to the opposite position on the right cushion. (The billiard table was, in fact, a metaphor chosen by Thomas Bayes in his famous essay, and Bayesian readers of *Significance* will no doubt permit themselves a smile of satisfaction at this point!) The blue lines show the various attempts. There is a great deal of scatter, but the arrival point is on average equal to the departure point (as regards vertical position, distance from 0). Thus the arrival point is on average as far up or down the table as its departure point. Reverse the question, however, and ask whether the departure point is on average equal to the arrival point as regards distance from 0, and the answer is "no". The departure point is on average closer to the middle of the table. The arrival points are much more spread out.

In fact, something like this may very well happen in science more generally. The iconoclast epidemiologist John Ioannidis has considered why it is that reported associations and effects generally turn out to be less impressive when studied again[5], a phenomenon that has notoriously affected gene expression as measured in micro-arrays but is by no means limited to that field. One explanation may simply be that the sort of shrinkage exhibited Figure 1 is occurring[6]. If the observed variation in expression is much greater than the true variation then shrinkage of that sort is inevitable.

> On average observed values are equal to true; but on average true values are not equal to observed. The words "on average" make all the difference

Is the Pope a Catholic?

"Yes", is the answer. Is a Catholic the Pope? "Almost certainly not", is the answer. Invalid inversion can lead to catastrophic errors of judgement. The error is obvious. Unfortunately, it is not so obvious that we avoid it, but training oneself to see it is a start.

References

1. Diaconis, P. and Freedman, D. (1981) The persistence of cognitive illusions. *Behavioral and Brain Sciences*, **4**, 333–334.
2. Good, I.J. (1995) When batterer turns murderer. *Nature*, **375**, 541.
3. Good, I.J. (1996) When batterer becomes murderer. *Nature*, **381**, 481.
4. Thompson, W. C. and Schumann, E. L. (1987) Interpretation of statistical evidence in criminal trials – the prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, **11**, 167–187.
5. Ioannidis, J. P. (2008) Why most discovered true associations are inflated. *Epidemiology*, **19**, 640–648.
6. Senn, S. (2008) Transposed conditionals, shrinkage, and direct and indirect unbiasedness. *Epidemiology*, **19**, 652–654; discussion 657–658.

Stephen Senn works for the Centre de Recherche Public de la Santé in Luxembourg. He is the author of *Dicing with Death* (Cambridge, 2003).