Probability, Justice, and the Risk of Wrongful Conviction

Jeffrey S. Rosenthal
University of Toronto, Canada

Abstract: We consider the issue of standards of proof in legal decisions from the point of view of probability. We compare ``balance of probabilities'' and ``beyond a reasonable doubt'' to the statistical use of p-values. We point out various fallacies which sometimes arise in legal reasoning. And we provide several examples of legal cases which involved probabilities, including some in which incorrect decisions were made and defendants were wrongfully convicted.

Keywords: probability, statistics, p-value, balance of probabilities, beyond a reasonable doubt, standard of proof.

Background and Context

I am a professor of statistics, and most of my work is fairly technical and mathematical (see www.probability.ca/jeff/research.html). But one day I wrote a book, *Struck by Lightning: The Curious World of Probabilities*, for the general public, which did unexpectedly well, especially in Canada. I was then interviewed by the media about such diverse topics as lottery probabilities, public opinion polls, violent crime rates, sports statistics, and more, and was even involved in probing a major lottery retailer fraud scandal involving millions of dollars and criminal convictions (see, for example, www.probability.ca/lotteryscandal). I was also invited to give talks to all sorts of different groups, from insurance brokers to financial investors, from humour therapists to gambling addiction counselors.

And then one day I was invited to speak to a prominent group of Canadian lawyers and judges. This invitation in turn led to others, and I ended up giving five different talks to five different groups of lawyers and judges (including the Irish Supreme Court and High Court justices) within a single year. This forced me to investigate the connection of probabilities and statistical analysis to the justice system, as I will now discuss.

Probability and Justice?

What is the connection of probability and statistics to justice issues? Well, both topics involve *evaluating evidence*, i.e., determining whether the available information is sufficient to draw certain conclusions. This perspective was nicely summarised by the James Bond villain Auric Goldfinger when he stated, “Once is happenstance. Twice is coincidence. The third time it’s enemy action.”

In probability and statistics, the possible conclusions might involve medical testing, or psychological analysis, or social science studies. We might try to determine if an observed difference is “statistically significant”, or if a certain probability is above some threshold, or if a corresponding p-value is less than some cutoff. In the justice system, the possible conclusions involve such questions as guilt or innocence. A court is then tasked with determining if a case has been proven “beyond a reasonable doubt”, or (for civil cases) by a “balance of probabilities” or a “preponderance of the evidence”. So what do these terms mean, and how are they related?

The justice term “balance of probabilities” seems closest to the world of probability and statistics. It seems to mean that a certain conclusion is more likely than not. So perhaps that means simply that the probability has to be more than 50%? Unfortunately even this connection is not so clear-cut. A classic example involves 1,000 people attending an event at which only 499 admission fares were collected. This means that a randomly-chosen attendee has probability 50.1% of not having paid admission. But surely no judge would award damages against a randomly-chosen attendee on that basis. Thus, even the simple-seeming “balance of probabilities” standard requires human judgement and cannot be interpreted in purely probabilistic terms.

The phrase “beyond a reasonable doubt” is even more challenging. It is generally agreed to mean something weaker than “certainty”, but something stronger than just “probably”. (For example, the Ireland Director of Public Prosecutions web site states: “The judge or jury has to be convinced beyond a reasonable doubt that a person is guilty. It is not enough for them to think that the accused is probably guilty.”) So does this mean the probability of guilt has to be more than 95%? more than 99%? more than 99.9%? Or that the corresponding p-value (i.e., the probability that we would have observed such evidence even if the accused were innocent) must be less than 5% or 1% or 0.1%? Once again, there is no clear standard, and the correspondence between probability/statistics and the standards of the justice system are hard to pin down.

Nevertheless, despite these challenges, it does seem that justice issues should be somewhat analysable in terms of probabilities and statistics. There are two major *risks* that need to be avoided: the risk of letting a guilty person go free, and (perhaps worse) the risk of wrongly convicting an innocent per- son. To explore these *risks* further, we next review some statistical practices, and then apply them to some specific legal cases.

How Statisticians Weigh Evidence

Consider a concrete example. Suppose your friend claims that she can distinguish Coke from Pepsi by taste alone. To test this claim, you pour Coke and Pepsi randomly into a series of glasses, and ask her to identify them. Suppose she identifies the first glass correctly (“Coke!”). This provides only slight evidence of her abilities, since she could have just gotten lucky. If she then also identifies a second glass correctly (“Pepsi!”), the evidence starts to increase. How many glasses in a row must she identify correctly before you would be convinced?

The classical statistical approach to this problem is to consider the *p- value*, i.e. the probability of observing such a result *if* your friend has no actual abilities and is just guessing randomly. So, if she guess right just once, then the p-value equals 1/2, or 50%. If she guesses right *twice* in a row, the p-value becomes (1/2) × (1/2) = 25%, where we multiply because (assuming random guessing) the different guesses are *independent*. If she guesses right *five* times in a row, the p-value equals (1/2) × (1/2) × (1/2) × (1/2) × (1/2) ≈ 3.1%.

Clearly, smaller and smaller p-values start to suggest that your friend’s guessing wasn’t just luck, but rather showed some true ability. But how small should the p-value be, to actually “prove” some conclusion? The usual standard, used throughout the medical and social sciences, is that a result is “statistically significant” if the p-value less than 5%, i.e. less than one chance in 20. Indeed, each time you take a medical drug, you are almost certainly consuming something which has been approved for treatment based on some study with some p-value which is less than 5%.

By this standard, if your friend guesses Coke versus Pepsi correctly twice in a row then this proves nothing, while if she guesses correctly five times in a row then this provides statistically significant evidence of her abilities. And, exactly the same reasoning applies to a new cure for a disease with a 50% fatality rate, which manages to save five patients in a row.

So far so good. But such statistical reasoning is not infallible. Sometimes the “evidence” is misleading due to incorrect reporting or biased sampling or incomplete recording. And even if the evidence is correct, sometimes the calculations or the conclusions are not.

One issue that arises is the *When To Multiply* question. That is, when is it appropriate for probabilities be *multiplied*? For example, if the probability of heads on one coin flip is 1/2, then the probability of heads on two coin flips is (1/2) × (1/2) = 1/4, because the coin flips are independent so multiplica- tion is valid. And in the above Coke/Pepsi example, if your friend is guessing without ability and the drinks are poured randomly, then each guess is independent of the next, so multiplication is again valid. But not always! For example, 49.2% of Americans are male, and 64% of Americans watch NFL football (according to a recent survey). So, does this mean the percentage of Americans who are male and watch NFL football is 49.2% × 64% = 31.5%? No, it’s actually 49.2% × 73% = 35.9%, since 73% of American males (and only 55% of American females) watch NFL football. That is, gender and football are not independent, so the multiplication is invalid and leads to too small a probability. I will consider this further below.

In addition, it is important to *interpret* p-values correctly. If your friend guesses Coke/Pepsi correctly five times in a row, then the p-value is 3.1%. This means that *if* your friend was guessing randomly, then the probability they would perform so well is 3.1%. This does *not* mean that the probability your friend was guessing randomly is only 3.1%. These two different probabilities are often conflated, which is sometimes called the *Prosecutor’s Fallacy*. In fact, the probability your friend was guessing randomly cannot be determined based on the experiment alone – it also depends on what you know or assume about your friend, and many other factors. In any case, it is *not* the same as the p-value, and indeed it could be very different.

Of even greater concern is *multiple testing*, or what I call the *Out Of How Many* principle. For example, in my book, *Struck by Lightning*, I tell the true story of running into my father’s cousin at Disney World, an event which seemed like one chance in hundreds of million. However, when you consider all the strangers I saw on that trip to Disney World, as well as all the people I *would* have been surprised to run into, it turns out that probability of running into *someone* surprising during a two-day trip to a crowded location is actually more like 0.5%, i.e. not so surprising after all.

In the context of the Coke/Pepsi experiment, if your friend keeps trying to guess all afternoon, and *eventually* guesses correctly five times in a row, then this proves nothing because they had so many chances to achieve this result.

Similarly, someone winning a lottery jackpot isn’t necessarily suspicious even though they have defied odds of one in tens of millions, because of all the *other* people who bought a lottery ticket and *could* have won instead. The same reasoning applies to all sorts of coincidences and suspicious – or surprising – seeming occurrences. That is, an apparently small p-value should always be treated with caution when other equally-surprising events were also possible.

The Prosecutor’s Fallacy, and the When To Multiply question, and the Out Of How Many principle, all have important applications to legal cases – as we now discuss.

A Legal Case: Sally Clark

Sally Clark was a solicitor in Cheshire, England. She had two sons, each of whom died in infancy with no apparent cause. The first death had been ruled a “cot death”, i.e., a case of Sudden Infant Death Syndrome (SIDS) in which babies sometimes suffocate without apparent explanation. But when the second infant also died, suspicions were raised, and Sally Clark found herself charged with double murder.

The prosecution case rested on probabilities. At her 1999 trial, the pae- diatrician Sir Roy Meadow testified that “the odds against two cot deaths in the same family are 73 million to one”. Clark was convicted on this basis, put in prison, and vilified in the press. She even had a third son temporarily removed from her custody. But was her conviction justified?

One issue is how the figure “73 million to one” should be interpreted. To the casual observer – or to the media, or even to a judge or juror – this might seem to be the probability that Clark is innocent. But the figure is actually a p-value, i.e. the probability that a law-abiding parent would have two infants die without apparent explanation, which is a rather different thing. Confusing the two is a classic example of the Prosecutor’s Fallacy!

A second issue is whether the figure “73 million to one” was computed correctly. Meadow obtained this figure by first saying that the probability of *one* child dying of SIDS was one chance in 8,543, and then saying that for *two* children we have to multiply to get a figure of $(1/8, 543) \times (1/8, 543) = 1/72, 982, 849 \approx 1/73, 000, 000$. However, this multiplication was not valid. Indeed, SIDS tends to run in families, so once a family has had one SIDS case, the second one is more likely – just like for male football watchers.

Furthermore, even the figure 1/8,543 for each individual SIDS death was not valid. The overall probability of SIDS in the U.K. had been estimated as 1/1,303. Meadow obtained 1/8,543 by “adjusting” for family circumstances that lower the SIDS probability (e.g. no smokers, someone employed, mother over 26 years old). But he neglected other factors which *raise* the probability (e.g. that SIDS is twice as likely for boys as for girls). So, for all of these reasons, the correct probability of two SIDS deaths in the same family was surely higher than the “one in 73 million” figure used in court.

But most important of all is the Out Of How Many principle. After all, there are tens of millions of families in U.K. alone. So, the fact that *one* of them had two SIDS deaths is not so surprising – much like the probability that *someone* wins the lottery jackpot. To convict solely on the basis of probabilities, the chances would have to be so low that we would *never* expect to see such an occurrence even once in any family in the U.K. or perhaps the entire world. By that standard, even a tiny p-value like “one in 73 million” is, well, not tiny enough. (By contrast, if Sally Clark was *already* under suspicion for some *other* reason, then a small p-value could well be convincing, since then the Out Of How Many principle might not apply.)

Sally Clark was convicted of double-homicide in November 1999, purely on the basis of Meadow’s probability calculation. However, statisticians soon noticed the flaws in the case. The venerable Royal Statistical Society noted (see: www.rss.org.uk/uploadedfiles/documentlibrary744.pdf) that Meadow’s approach was “statistically invalid”, and declared that “The case of R v. Sally Clark is one example of a medical expert witness making a serious statistical error, one which may have had a profound effect on the outcome of the case.” For these and other reasons, Clark was ultimately acquitted on her second appeal, after more than three years in jail. But she never recovered psychologically, and died of alcohol poisoning four years later.

Meanwhile, the U.K. General Medical Council ruled that Meadow's evidence was "misleading and incorrect", and that he was guilty of "serious professional misconduct". He was effectively barred from any future court work. Furthermore, the prosecution pathologist Alan Williams was found to have not reported evidence about an *infection* in the second son (which may have suggest death by natural causes). The GMC found him, too, guilty of serious professional misconduct. As a result, several other similar convictions were also overturned on appeal. A valuable lesson had been learned.

An Earlier Case: Malcolm Collins

On June 18, 1964, in Los Angeles, an elderly lady was pushed down in an alley, and her purse was stolen. Witnesses said: a young Caucasian woman, with a dark blond ponytail, ran away with the purse, into a yellow car, which was driven by a Black man, who had a beard and moustache. Four days later, Malcolm and Janet Collins were arrested, primarily because they fit these same characteristics (at least mostly – Janet's hair was apparently light blond rather than dark blond).

At trial, the prosecutor called "a mathematics instructor at a nearby state college" (whose identity no one seems to know). The prosecutor told the mathematics instructor to assume certain "conservative" probabilities:

- Black man with a beard: 1 out of 10
- Man with moustache: 1 out of 4
- White woman with blond hair: 1 out of 3
- Woman with a ponytail: 1 out of 10
- Interracial couple in car: 1 out of 1,000
- Yellow car: 1 out of 10

The mathematics instructor then computed the probability that a random couple would satisfy all of these criteria, by – you guessed it – multiplying:

$$(1/10)\times(1/4)\times(1/3)\times(1/10)\times(1/1000)\times(1/10) = 1/12{,}000{,}000$$

It was thus asserted that there was just one chance in 12 million that a couple would have these same characteristics if they were not guilty. Malcolm Collins was convicted at trial, primarily based on this "one in 12 million" probability.

Was this probability calculation valid? Surely not. For one thing, those individual probabilities were just *assumed*, without evidence. Furthermore, the multiplication was again invalid: for example, most men who have beards also have moustaches, so (like the male football watchers) these factors are surely not independent. (And, if you have a Black man and a White woman, then *of course* you have an interracial couple! Perhaps the prosecutor may have meant that one in 10 Black men have beards, not that one man in 10 is Black with a beard, but the interpretation is rather difficult to sort out.) So, the asserted probability of "one in 12 million" is highly questionable.

Even more important, once again, is the Out Of How Many principle. Los Angeles County in 1964 had a population of 6,537,000, and thus approximately one million couples (which could termed the "suspect population"). So, even with odds as small as one in 12 million, the probability of there being two such couples is actually fairly large. To convict Malcolm Collins on the basis of probabilities alone seems inappropriate in this case.

The case of Malcolm Collins eventually made its way to the Supreme Court of California. Their 1968 judgment (found here: scholar.google.com/scholarcase?case=2393563144534950884) began, “We deal here with the novel question whether evidence of mathematical probability has been properly introduced and used”. They rightly observed that “the testimony as to mathematical probability infected the case with fatal error”. However, they then further insisted that the trial’s probability calculations had “distorted the jury’s traditional role of determining guilt or innocence according to long-settled rules”, concluding that “Mathematics, a veritable sorcerer in our computerized society[1], while assisting the trier of fact in the search for truth, must not cast a spell over him. We conclude that on the record before us defendant should not have had his guilt determined by the odds”. They overturned the conviction on that basis. Now, I am glad that the conviction was overturned, due to the flaws in the probabilistic reasoning. But I wish they hadn’t implied that guilt should *never* be determined by the odds – I disagree and think that is going too far.

Another Case: Lucia de Berk

Lucia de Berk was a nurse who worked on three different hospital wards in The Hague, Netherlands. She was arrested after it was discovered that she was on duty for 14 of 27 “incidents” (i.e. patient deaths or near-deaths) in her three wards (51.9%), despite working just 203 of the 2,694 shifts in her three wards (7.5%). At her trial, the prosecution asserted that there was just one chance in 342 million that such an imbalance would occur by chance alone. de Berk was convicted of multiple murders and attempted murders in March 2003, primarily on the basis of this “1 in 342 million” probability. Was her conviction justified?

A first question is whether the evidence (i.e. facts) were accurate. There was some controversy about whether all of these incidents had actually taken place *during* de Berk’s shifts, as opposed to just before or just afterwards. Furthermore, the definition of “near-death” might have been adjusted *post hoc* to include more incidents during de Berk’s shifts. Related to this, de Berk may have been assigned to extra elderly/terminal patients due to her experience as a nurse, which may have provided an alternative explanation of any excess number of incidents. These issues were all debated vigorously following her conviction.

In addition to the above, many of our previous concerns apply. Once again, the Prosecutor’s Fallacy must be avoided: the probability that de Berk is guilty *given* the observed facts is quite different from the probability that the observed facts *would* have arisen if she were innocent. Even more important is the Out Of How Many principle. The prosecution statistician, Henk Elffers, had tried to account for this by multiplying by 27 (the number of nurses in one of the hospitals), but arguably he should really have multiplied by the number of nurses in the entire Netherlands or even the whole world.

After de Berk’s conviction, various statisticians objected. In particular, four Dutch statisticians alluded to the Out Of How Many principle by saying “the data . . . is used twice: first to identify the suspect, and after that again in the computations of Elffers’ probabilities” (Meester, Collins, Gill & Lambalgen, 2006). They made numerous “adjustments”, and eventually increased the p-value from “1 in 342 million” to 0.022 (i.e. 1 chance in 45), a p-value which is surely too large for conviction.

de Berk’s convictions were upheld on first appeal in 2004. However, enough doubts had been raised about the probability calculations that the conviction was upheld primarily on *other* grounds, notably elevated digoxin levels in some of the corpses (which could be evidence of poisoning).

[1] If they thought society was “computerized” in 1968, what would they think today?

However, the hypothesis of digoxin poisoning was disproven by 2007, leading to the case being reopened in 2008, and a not guilty verdict being delivered on second appeal in 2010. Lucia de Berk is now a free woman.

Of course, none of this precludes the possibility that Lucia de Berk might have been guilty. For example, she may have killed some terminal patients out of *mercy*, to relieve their suffering in their final days. Indeed, on the day of one of her elderly patient's death, de Berk wrote in her diary that she had "given in to her compulsion" (though she later claimed she was referring to her compulsion to read Tarot cards). While this fact was introduced at her trial, her conviction was based primarily on the statistical evidence. And, as we have seen, the statistical evidence wasn't sufficiently convincing.

Discussion

I have presented three different legal cases where people were convicted of serious crimes primarily on the basis of faulty probability calculations. It may be tempting for some people to conclude from this – as the Supreme Court of California perhaps did in 1968 – that probabilities should never be used to convict anyone of anything.

I think that this is going too far, and that statistical analysis *can* sometimes help to achieve justice after all, provided that it is used with caution. One example of this is the lottery retailer scandal mentioned earlier. In that case, I was able to determine that lottery retailer ticket sellers had won more major lottery prizes than could be reasonably explained by chance alone. This conclusion became a huge news story in Canada, and led to millions of dollars in lottery repayments, and several criminal convictions for fraud (see www.probability.ca/lotteryscandal). This illustrates how careful statistical calculations which take into account the factors mentioned above can identify criminal activity and achieve justice.

An interesting related story is that of Waneta & Tim Hoyt. They had five babies in New York State during 1965 – 1971, *all* of whom died in infancy (at 3, 28, 1.5, 2.5, and 2.5 months old, respectively). The deaths were all identified as SIDS, and indeed a pediatrician used them to publish a scholarly article about SIDS' strong genetic linkage (Steinschneider, 1972). Apparently no foul play was suspected, and in fact the Hoyts were later allowed to adopt a son (who survived to adulthood) in 1977. Years later, in 1985, some prosecutors and pathologists became suspicious, and investigated. Eventually, Waneta Hoyt confessed to suffocating all five of the children, to stop them from crying. She later "recanted" her confession, but was nevertheless convicted in 1985 of five murders; she died in prison in 1998 at the age of 52. It would appear, at least in hindsight, that her murderous ways should have been detected much sooner – but instead the genetic linkage was believed so strongly that even five deaths were not considered suspicious.

As these examples illustrate, it is difficult to decide when probabilistic evidence is sufficient to justify a criminal conviction. The calculation and interpretation of p-values is often challenging. Overly aggressive or simplistic calculations run the risk of convicting innocent people, while overly cautious analyses run the risk of setting guilty parties free. Alternatively, it is possible to take a Bayesian approach to this question (see, for example, Meester et al., 2006, section 6), but that requires specifying prior probabilities which is itself problematic and subjective.

Nevertheless, I do believe that probabilities can and should be used in criminal trials (among other places). Such probabilities must be carefully computed, accounting for such issues as the accuracy of the data, the When To Multiply question, the Prosecutor's Fallacy, and (perhaps most important of all) the Out Of How Many principle. If all of these factors are carefully taken into account, then probabilities can indeed be used to draw conclusions and avoid risks, even about criminal activity.

References

Meester, R., Collins, M., Gill, R., & van Lambalgen, M. (2006). On the (ab)use of statistics in the legal case against the nurse Lucia de B. Law, *Probability and Risk, 5*, 233–250.

Steinschneider, A. (1972). Prolonged apnea and the sudden infant death syndrome: clinical and laboratory observations. *Pediatrics, 50*(4), 646–654.