Recognising and Avoiding Fallacies in Interpreting Statistical Evidence

Henry Prakken
Faculty of Law, University of Groningen &
Faculty of Science, Department of Information and Computing Sciences, Utrecht University
The Netherlands

19 February 2021

Abstract

Legal fact finders are increasingly confronted with statistical evidence presented by forensic or other experts. Various studies have revealed that those involved in court cases find it often very hard to interpret such evidence. This chapter discusses how basic knowledge of standard and Bayesian probability can help recognising and avoiding some of the most frequently occurring fallacies when interpreting statistical evidence.

1 Why should legal professionals and law students know about probability theory?

When fact finders (judges or juries) have to determine the facts of a case, they are increasingly confronted with evidence that is presented in terms of probabilities (*statistical evidence* for short). Often this concerns forensic trace evidence, such as DNA evidence, fingerprint evidence, footprint evidence or tire tracks evidence. Such evidence usually comes with a so-called *random-match probability*, which is the probability that a potential source of the trace matches with the trace by coincidence, that is, if the potential source is not the actual source of the trace. Such probabilities are often based on statistical information concerning the frequencies of occurrences of particular patterns in a population, such as the frequency of specific DNA patterns or fingerprint patterns or feet shapes or car tire profiles. Other examples of statistical evidence are when experts report on the relative probability of finding a particular piece of evidence given alternative hypotheses about what may have happened. For instance, medical experts may report that they regard the injuries of a particular child much more probable if they are caused by child abuse than if they are caused by an accident. Such probabilities are called *likelihood ratios*. They can be based on statistics or on the expert's expertise. Sometimes an expert reports only half of a likelihood ratio, such as a medical expert who asserts that that the probability that two babies in the same family die of cot death is one in 73 million.

It turns out that those involved in court cases find it often hard to interpret statistical evidence. This is no surprise since research in cognitive science has provided ample evidence that people are in general poor in dealing with probabilities (Tversky & Kahneman 1974; Kahneman 2011). While in daily life this may not be a big problem, in court this can result in serious miscarriages of justice. For example, in the UK Sally Clark was in 1999 convicted for having murdered her two babies, who

had both suddenly died when only their mother was at home. In 2003 she was acquitted in a second appeal after eminent statisticians had argued that the medical expert in this case, who had estimated the probability that two babies in the same family of the type of the Clarkes die of unexplained natural causes is one in 73 million, had made serious statistical errors. Similar miscarriages of justice have happened in other countries. For example, in the Netherlands the nurse Lucia de Berk was initially convicted of murdering seven young children who had all died while being at an intensive care unit when Lucia de Berk was on duty. Later she was acquitted in a revision case, which was opened after statistical experts had shown that an initial expert had made serious errors when he estimated the probability that seven children at an ICU die of natural causes while the same nurse is on duty is one in 342 million. Moreover, the experts convincingly argued that the judges and prosecutors in the case had dramatically misinterpreted the experts' probability estimate (Meester et al. 2006; Derksen & Meijsing 2009). In both the UK and The Netherlands these cases gave rise to fierce debates on the use of so-called Bayesian probability theory in court. In the USA a similar debate had arisen much earlier, after in 1968 a couple had been convicted of a robbery partly on the basis of statistical evidence offered by a university professor. Here too, other experts showed that the professor had made serious statistical errors (Tribe 1971; Lempert 1986).

One theme in these debates is how the police, juries, judges, prosecutors and others involved in criminal investigation and (criminal or civil) trials can be safeguarded against reasoning errors when interpreting statistical evidence. This question is practically very important given the increasing amounts of statistical evidence presented in court (a development that started with the rise of DNA evidence). Another theme in this debate is of a more theoretical nature, namely, the question what is a good theory of rational reasoning about evidence. Should fact finders ideally think in terms of probability theory or can other modes of reasoning (for instance, argumentation- or scenario-based) also be rational? This chapter is devoted to the first, practical question only, except for a few remarks in the concluding section on the theoretical issue. For a recent collection of papers on theories of rational legal proof see Prakken et al. (2020). The main aim of the present chapter is to explain the basics of standard and Bayesian probability theory and to illustrate how it can be used to recognise and avoid some of the most frequent reasoning errors made in court with statistical evidence.

To fulfil this aim, I shall first present some examples involving types of statistical evidence that are often misinterpreted (Section 2), after which in Section 3 I introduce the basics of probability theory. In Section 4 I use these basics to discuss some statistical fallacies based on inverting conditional probabilities and ignoring base rates. Analysing other fallacies requires the use of *Bayes's Theorem*, which I will present and apply in Section 5. Then I shall in Section 6 discuss some issues surrounding the use of probability theory in court. I conclude in Section 7 with a practical recommendation for legal fact finders and some issues for research for legal scholars.

2 Examples of statistical evidence in legal evidential reasoning

In this section I give some typical examples of statistical evidence presented in court cases. The first three examples are artificial but still realistic while the other four examples are real.

Drugs test Suppose Bob is involved in a car accident and is suspected of having used a kind of drug. He is subjected to a drugs test that is known to be 99% reliable, that is, 99% of the drug users are identified as drug users and 99% of those who did not use the drug are identified as not having used the drug. Bob tests positive. The judge may want to know what is the probability that Bob has used the drug given the positive test. Many people are inclined to say '99%' but we shall see that the example does not contain enough information to answer the judge's question.

Paternity test The following example is half real, half artificial. The Dutch company Verilabs, which offers paternity DNA tests claims at its website www.dnavaderschapstest.nl that

> "In a paternity test, Verilabs shows with a confidence of more than 99.99% whether a man is the biological father of a child" (my translation, HP).

Imagine that Mary claims that John is the father of her child and John tests positive in a Verilabs test. The judge having to adjudicate Mary's claim will want to know what is the probability that John is the father of Mary's child given the positive test. Many will be inclined to say 'more than 99.99%' but we shall see that this example, too, does not contain enough information to answer the judge's question.

Blue and green taxis The third example was constructed by Tversky & Kahneman (1974) for a famous experiment on how well people interpret probability information. On a misty winter night a taxi hits another car and disappears in the night. A witness says that he saw that the taxi was blue. In the town where the accident happened that are two taxi companies, which together own 100 taxis. One company owns 85 of the taxis which are all green, while the other company owns the remaining 15 of the taxis, which are all blue. The witness is tested on his reliability and turns out to correctly report the colour of 80% of the taxis that are shown to him. The test subjects had to answer the question what is the probability that the taxi that hit the other car is blue given the testimony. Many (though not all) answered '80%' but we shall see that this is a fallacy.

Sally Clark The next example is a tragic case that happened in England (below I follow the description in Dawid 2005, Section 4.3). In December 1996 Sally Clark's first son suddenly died, 2,5 months old, while he was alone at home with his mother. In January 1998 Sally's second son died, 2 months old, also while being at home alone with his mother. Sally was accused of having killed her sons but Sally claimed they had died of natural causes (maybe cot death). A paediatrician estimated the probability that one child dies from unexplained natural causes in a family such as the Clarks is 1 in 8500. He then multiplied this probability with itself to conclude that the probability that two children die from unexplained natural causes in a family such as the Clarks is 1 in 73 million. Many may be tempted to infer from this that Sally almost certainly killed her two sons and indeed the jury found Sally Clark guilty and her first appeal was dismissed. However, we shall see that this inference is based on at least two reasoning errors.

Tire tracks The next example is a Dutch criminal case from 2014[1] in which tire tracks were found at the crime scene that match with the tire profile of the car owned by the accused. An expert witness testified that

> "The probability that a random Dutch car (…) has tire profiles that match with the observed tire tracks is (…) 1 in 5000" (my translation, HP)

The court of appeal interpreted the expert's testimony as follows:

> The probability that the tire tracks are caused by a random other car is (…) estimated at 1 in 5000, which yields a high probability that the accused's car made the tire tracks." (my translation, HP)

While many may be tempted to agree with the court of appeal, we shall see below that the court's inference from the expert's testimony is based on a reasoning fallacy.

Child abuse In a Dutch child abuse case in 2014[2] a medical expert testified that

> "The combination of the brain injury, the bruises under the hard meninges, the retinal haemorrhages and the skin lesions is very much more probable in a non-accidental event than in an accidental event or medical condition" (Google Translate's translation, HP)

The criminal court added that the expert witness had stated that

> "there is no higher degree of probability in this field than 'very much more probable'." (my translation, HP)

The court then concluded as follows:

> "The court accepts the conclusions of expert (…) and infers from this that the victim must have been seriously assaulted (…)". (my translation, HP)

We shall see that the court's inference is based on the same fallacy as in the tire tracks case. Note that the expert did not report numerical probabilities; I shall explain later in Section 5 that this does not prevent the application of probability theory to the example.

Denis Adams The final example concerns a rape case in England in the early 1990s (My discussion of this example is based on Dawid's 2005 discussion of the same example). In 1991 a woman was raped in Hemel Hempstead near London. In 1993 Denis Adams was arrested for another offence and his DNA profile was discovered to match with the DNA profile of a semen sample obtained from the rape victim. He was then accused of being the rapist. (From the discussion in Dawid (2005) it seems that

[1] ECLI:NL:RBNHO:2014:10689

[2] ECLI:NL:RBZWB:2014:4249

during the case it was uncontested that the semen sample was from the rapist and that the only issue was whether Adams was the source of the semen sample and therefore guilty of rape. Therefore, I will in my discussions of this example equate the issues of identification and guilt; note that in other cases such an equation may not be justified.) A prosecution's forensic expert estimated the probability that a random person's DNA would match with the DNA found at the crime scene as 1 in 200 million. Such a probability is called a *random-match probability*. The defence made a lower estimate of this probability of 1 in 2 million. Besides the DNA evidence there were two further pieces of evidence in the case. First, the victim had failed to recognise Adams in a lineup and, second, Adam's girlfriend testified that he had spent the night of the rape with her. The jury found Adams guilty of the rape.

Below I shall use the Adams case to once more illustrate that statistical DNA evidence has to be interpreted with care, but also to illustrate that DNA evidence has to be combined with other evidence if available.

3 The basics of probability theory

In this section I introduce the basics of probability theory step-by-step. Along the way I will return to several of the examples from Section 3.

3.1 Basic properties of probabilities

In probability theory, uncertainty concerning the truth of a statement is expressed in a number between 0 and 1 (or equivalently between 0% and 100%). A probability of 1 (or 100%) means that the statement is certainly true, a probability of 0 (or 0%) means that it is certainly false, and every number or percentage between these extremes expresses a degree of uncertainty. An important property of probabilities is that if two statements cannot be true at the same time and together exhaust all possibilities, then they add up to 1 (or to 100%). For example, if the probability that it will rain tomorrow is 0.8 (or 80%) then the probability that it will not rain tomorrow is 0.2 (or 20%). This property holds for a statement and its logical negation, as in the just-given example or for 'John was at the crime scene' versus 'John was not at the crime scene' but it also holds for incompatible statements that leave no other possibility given the way the world is (for instance, 'the taxi that hit the other car was green' and 'the taxi that hit the other car was blue' if we are sure that only a green or blue taxi could have caused the accident).

3.2 Statistical independence

An important concept in probability theory is statistical independence. Two statements A and B are *statistically independent* of each other if information about the probability that A is true is irrelevant for determining the probability that B is true, otherwise they are *statistically dependent*. For instance, let A be 'I flip a coin the first time' and B 'I flip the coin a second time'. Assuming the coin is fair, the probability that I flip heads the first time is ½ (that is, 50%). Clearly, the probability that I flip heads the second time also is ½ (50%) since the two coin-flipping events do not influence each other. When two statements A and B are statistically independent, then the probability that they are both true ('A and B') can be calculated by multiplying the

probabilities that each of them is true. So the probability that I flip heads both the first and the second time equals ½ x ½ = ¼ (or 25%). However, when A and B are statistically dependent of each other, then this calculation is invalid and the probability of A and B has to be determined independently from the probabilities of A and B. For example, when I throw a fair dice once, and A = "I throw an even number" and B = "I throw a 6", then the probability of A equals ½ (since there are three even and three odd numbers) while the probability of B equals 1/6. Clearly, if we are told that I threw an even number, this influences the probability that I threw a 6, since now there are only three possibilities 2, 4 and 6, so the probability that I threw a 6 has increased to 1/3. This is unequal to the product of the two probabilities of A and B, which equals ½ x 1/6 = 1/12.

While in simple chance games it is easy to see whether two statements are statistically independent, in the law this is often different. Consider again the **Sally Clar**k case. We saw that the paediatrician multiplied the probability of 1 in 8500 that one child dies from unexplained natural causes in a family such as the Clarks with itself to obtain that two children die from unexplained natural causes in a family such as the Clarks. This calculation results in 1 in 73 million, which is such a small probability that it made many believe that Sally Clark almost certainly killed her two sons. Later we shall identify this as another fallacy but for now we can identify the first fallacy in the paediatrician's reasoning, namely, to assume without justification that the two deaths of Sally Clarks sons are statistically independent. Other experts pointed out that these two deaths may be related because of shared genetic, social or domestic characteristics, so the paediatrician should not have multiplied the probabilities of the single deaths. The other experts, taking the dependencies into account, estimated the probability that two babies die of natural causes in a family such as the Clarks as 1 in 850,000. This probability is still very small so many may still be have been inclined to infer from this that Sally Clark almost certainly killed her two sons. However, in Section 5 I shall show that this inference is fallacious.

3.3 Conditional probabilities

Another important concept in probability theory is that of a conditional probability. Such a probability expresses the probability that some statement is true given that, or assuming that some other statement is true. Here are some examples:

> the probability that I throw a 6 with a fair dice given that I throw an even number;
> the probability that FC Barcelona will win next year's Champions Leave given that Lionel Messi leaves the club after this season;
> the probability that the suspect was at the crime scene given that a witness says he saw him a the crime scene;
> the probability that the suspect committed the crime as charged given the available evidence.

In evidential reasoning conditional probabilities are crucial, since in the end we are interested in the probability that some legally relevant statement is true given the available evidence. Several probabilities in our examples from Section 2 are conditional. For example:

> the probability that Bob tests positive in a drug test given that Bob uses the drug;

the probability that Bob uses the drug given that he tests positive in a drug test;
the probability that John is the father of Mary's child given the positive DNA paternity test;
the probability that a random car's tire profile matches with the tire tracks found at the crime scene given that the car did not make the tire tracks;
the probability that this child has these injuries given that they were caused by an accidental event or medical condition;
the probability that the child was seriously assaulted given that it has these injuries.

Below we will see that that the examples contain more conditional probabilities, although this is not always easy to recognise from the natural language used by experts or judges. It will turn out that the vagueness of natural language is one reason why probabilities reported by an expert are often misinterpreted.

Conditional probabilities have much the same properties as unconditional probabilities. They are also between 0 and 1 (or between 0% and 100%), and when two statements A and B exclude each other and are jointly exhaustive of all possibilities, then their probabilities given the same condition C add up to 1 (or to 100%). For example, the probability that Albert was at the crime scene given that Carole says she saw him and the probability that Albert was not at the crime scene given that Carole says she saw him add up to 1 (or to 100%). So if we estimate the first of these probabilities as ¾ (or 75%) then the latter probability is ¼ (or 25%). We will need this property of conditional probabilities later to derive useful conclusions from Bayes's theorem.

Conditional probabilities allow a precise definition of when two statements A and B are statistically independent of each other. We say that statement A is statistically independent of statement B if the conditional probability of A given B equals the unconditional probability of A. In other words, whether B is true or not does not matter for the probability that A is true.

This new definition of statistical independence is a good way to test whether two statements are statistically independent. For example, we now have another way to express that the paediatrician in the **Sally Clark** case made an unjustified statistical independence assumption: the unconditional probability that a child dies from unexplained natural causes in a family as the Clarks is not the same as the conditional probability that a child dies from natural causes in a family as the Clarks given that another child in the same family has died from unexplained natural causes.

4 Inverted conditional probabilities and the base rate fallacy

Even without introducing Bayes's Theorem we can give a systematic account of some types of statistical fallacies. In legal evidential reasoning one cause of such fallacies is that conditional probabilities are often erroneously inverted. For example, in our **drugs use** example I said that many people are tempted to conclude that the probability that Bob used the drug given the drugs test is 99% since the drugs test is 99% reliable. However, this confuses the following two conditional probabilities:

the probability that a person is tested positive given that he used the drug;
the probability that a person used the drug given that he is tested positive

The reliability (often called accuracy) of a drugs test does not equate the second but the first conditional probability (together with the probability that the person is tested negative given that he did not use the drug). Recall that in the example we said that 99% reliable means that 99% of the drug users are identified as drug users and 99% of those who did not use the drug are identified as not having used the drug. These two frequencies correspond to the following two conditional probabilities:

> the probability that a person is tested positive given that he used the drug is 99%;
> the probability that a person is tested negative given that he did not use the drug is 99%.

This is not the same as saying

> the probability that a person used the drug given that he is tested positive is 99%;
> the probability that a person did not use the drug given that he is tested negative is 99%.

Now the crucial thing is that to make the step from the probability that Bob is tested positive (negative) given that he used (did not use) the drug to the probability that Bob used (did not use) the drug given that he is tested positive (negative) we need more information. The information we need is how many people in the population we are considering use the drug (called the *base rate* of drug users). To see this, assume that it is known that 0.5% of the population uses the drug. For ease of calculation, let us consider a population of 100,000 people. We then know that 500 of them use the drug while the remaining 99,500 do not use it. Of the 500 people who use the drug, 99%, so 495 people, will correctly test positive while 1%, so 5 people, will incorrectly test negative (note that to calculate the number of negatively tested people we use the property that the probabilities of a statement and its negation add up to 1 (or 100%). In addition, of the 99,500 people who do not use the drug, 99%, so 98,505 people, will correctly test negative but the remaining 1%, so 995 people, will incorrectly test positive. So of all the 495+995=1490 people who will test positive (including Bob), only 495/1490, which approximately equals 33.2%, use the drug. Bob could be any of these 1490 people who tested positive, so given the positive test, the probability that he used the drug is only 33.2%, even though the drug test is 99% reliable.

This calculation is summarised in the following table:

Persons	**Total**	**Positive**	**Negative**
Drug users	500	495	495
Non drug-users	99,500	995	98,505
Total population	100,000	1490	98,510

Table 1: the drug use example

Since Bob is tested positive, he is somewhere in the ‘Positive’ column but we don’t know whether he is in the ‘Drug users’ row or in the ‘Non drug-users’ row. Then the probability that he is in the ‘Drug users’ row is 495/1490 = 33.2%, as we calculated.

The reason for this at first sight counterintuitive outcome is that there are so many more non drug-users than drug users that even with a highly reliable test more non-

users will be incorrectly tested positive (the so-called *false positives*) than that users will be correctly tested positive (the so-called *true positives*). The failure to see this is often called the *base rate fallacy* (after Tversky & Kahneman 1974), which (in our example) is the fallacy to neglect the high base rate of non drug-users in the population.

We now see that there may be some truth in the claim of many athletes tested positive in a doping test that they are innocent: if the great majority of the athletes does not use the drug for which the athlete is tested, then even with a highly reliable doping test there may be more false positives than true positives.

The **Blue and green taxis** example can be analysed in the same way. Recall that a witness is known to correctly report the colour of 80% of the taxis show to him, that the witness said that the taxi that hit the other car was blue, and that in the town of the accident 85 of the taxis are green while the remaining 15 are blue. We said that many are tempted to infer that the taxi is blue given that the witness says so is 80%. However, this inference is another instance of the base rate fallacy, since it ignores that there are so many more green taxis than blue taxis in town that the probability of a false positive identification by the witness is still high even though he is 80% reliable. The following table confirms this:

Taxis	**Total**	**"Blue"**	**"Green"**
Blue	15	12	3
Green	85	17	68
Total taxis	100	29	71

Table 2: the blue and green taxis example

We see that the witness will correctly identify 12 of the 15 blue taxis as blue and incorrectly identify 3 blue taxis as green. Likewise, he will correctly identify 68 of the 85 green taxis as green and incorrectly identify 17 green taxis as blue. So only 12 of the 29 taxis he will identify as blue are in fact blue, so given that he identified the taxi that hit the other car as blue, the probability that it is indeed blue is 12/29, which approximately equals 41.4%. The reason for this low probability despite the witness's high reliability is that there are many more green taxis than blue ones, which makes a false positive more probable than a true positive.

We now know which information was lacking in the **Paternity test** example, in which John was tested positive in a DNA paternity test to see whether he was the father of Mary's child. What we also need to know is how many men could be the father of Mary's child. Unlike in the previous example, this is not so easy to express in population frequencies; we have to make an estimate on other grounds, perhaps based on specific evidence about John and Mary or simply on the basis of our commonsense, given where Mary lives and how many male adults live in her area. Let us first assume that there are 100,000 potential fathers; in practice it may be hard to estimate such a precise number, but we need it to explain the correct reasoning; with less precise numbers the reasoning stays the same although the input of the reasoning may be less than certain.

Recall also that Verilabs claimed that Verilabs claimed that its test shows with a confidence of more than 99.99% whether a man is the biological father of a child. We now know that Verilabs cannot have meant the following probabilities

> the probability that John is the father given a positive test is more than 99.99%;
> the probability that John is not the father given a negative test is more than 99.99%

since this simply not the information that is known about a medical test; what is always reported is the probability of a test outcome given a hypothesis, not the probability of a hypothesis given a test outcome. So what Verilabs means is

> the probability that John tests positive given that he is the father is more than 99.99%;
> the probability that John tests negative given that he is not the father is more than 99.99%.

Let us apply these probabilities to our guess that there are 100,000 potential fathers of Mary's child, for convenience ignoring the 'more than'. Of these 100,000 potential fathers, one is the real father while the other 99,999 are not the real father. The real father will (with a very small error margin of 0.01% that we can safely ignore) certainly test positive) while of the other 99,999 men 99.99% will test negative, which are 99,989 men, while still 0.01% of them will test positive, which are 10 men. Note that in the latter calculation we apply the property that the conditional probabilities of a statement and its negation given the same condition add up to 100%: since the probability of a negative test given that the person is not the father equals 99.99%, the probability of a positive test given that the tested person is not the father equals 100% - 99.99% = 0.01%. Table 3 summarises our analysis:

Potential fathers	Total	Positive	Negative
Real father	1	1	0
Other men	99,999	10	99,989
Total men	100,000	11	99,989

Table 3: paternity test example

Only one of the 11 men who test positively is the real father. Since John tested positive we know that he is one of these 11 men but we do not know which one. So the probability that he is the real father given the positive test is 1/11, which approximately equals 9.1%.

Does this mean that a DNA paternity test is weak evidence after all? No, since if we succeed in reducing the number of potential fathers with other evidence, then the positive test quickly makes it probable or even highly probable that John is the father. For example, with 10,000 potential fathers there will be only one false positive test (0.01% x 10,000), so then the probability that John is the father given the positive test is ½ = 50%. And with 1000 potential fathers the probability of a true positive is 10 times greater than the probability of a false positive test, which gives a probability of approximately 91% that John is the father. Finally, assume that John admits that he had sexual intercourse with Mary 9 months before the child was born and that we have reason to believe that around the same time Mary had sexual intercourse with 9 other men, so the number of potential fathers is 10. Then the probability of a true positive is 1000 times greater than the probability of a false positive, which yields a

probability of 99.9% that John is the father given the positive test (that is, if John cannot provide any evidence against his fatherhood; see further Section 5 below). The last two calculations are hard to display in a table, since the expected number of false positive tests is smaller than 1. In Section 5 we will see that Bayes's Theorem allows us to make the precise calculations.

The final example of Section 3 that can be analysed with this section's tabular method is the **tire tracks** example, in which car tire tracks found at the crime scene matched with the tire profile of the accused. Recall that an expert witness testified that

> "The probability that a random Dutch car (…) has tire profiles that match with the observed tire tracks is (…) 1 in 5000"

The court of appeal interpreted the expert's testimony as follows:

> The probability that the tire tracks are caused by a random other car is (…) estimated at 1 in 5000, which yields a high probability that the accused's car made the tire tracks."

By now the reader will have understood that in order to draw useful inferences from the expert's probability, we must know the number of Dutch cars. Let us assume that there are 5 million Dutch cars; only one of them made the tire tracks found at the crime scene (for simplicity we ignore the possibility that a foreign car made the tracks). The tire profile of that car will surely match with the tire tracks found at the crime scene but also those of 1 in 5000 of the other Dutch cars will match, which are 1000 cars. The accused's car could be any of these 1001 cars, so the probability that his car caused the tire tracks given the match is just 1/1001, which approximately equals 0,1%. This is a far cry from the court's conclusion to "a high probability that the accused's car made the tire tracks".

What has gone wrong here? It may be that the court was misled by the ambiguous nature of the expert's statement, which does not have a clear conditional structure like "the probability of this given that is …". This happens more often: natural language is not as precise as mathematical language, which is one reason why expert testimonies on probabilities are so often misinterpreted. Another reason is that judges and prosecutors usually are not properly trained in probability theory and its use in evidential reasoning.

While the tabular method we used in this section is suitable for several types of examples, especially those in which probabilities can be based on frequencies of (human or other) populations, this does not hold for all types of examples. First, the tabular method does not apply well to examples with multiple pieces of evidence, as in the Denis Adams example. Furthermore, not all probabilities can be based on frequencies; sometimes probabilities pertain to specific events, such as the probabilities in the child abuse example that the observed injuries are caused by an accident. Finally, the way the expert in the child abuse example reported the probabilities, namely, as a ratio between two probabilities, cannot easily be explained with the tabular method. All three kinds of examples are better analysed with Bayes's Theorem, of which the tabular method is but a special case.

5 Bayes's Theorem and its use in legal evidential reasoning

In this section I present Bayes's Theorem, a theorem that can be mathematically derived from the axioms of probability theory. Bayes' Theorem is at the heart of the so-called Bayesian way of applying probability theory. After presenting the theorem, I shall first apply it to some examples from the previous section, to illustrate that the tabular method used in that section is equivalent to a special case of Bayes's Theorem. Then I shall apply the theorem to the remaining examples from Section 2.

Bayes's Theorem is about the relation between two mutually exclusive hypotheses and evidence pertaining to these hypotheses. In legal evidential reasoning the hypotheses can be any factual statement the truth of which is to be determined in court, such as 'John is/is not the father of Mary's child', 'Bob used/did not use drug X', The taxi that hit the other car was blue/green', 'Sally Clark killed her two suns/Sally Clarks two suns died of unexplained natural causes', and so on. There are several mathematically equivalent formulations of Bayes's Theorem. I will present the so-called odds version, which states a mathematical relation between three ratios of probabilities. Consider two hypotheses H1 and H2 and E a piece of evidence pertaining to H1 and H2 (later I will instantiate these symbols with specific examples). Then the *prior odds* states the ratio between the unconditional probabilities of H1 and H2:

$$\frac{\text{The probability of hypothesis H1}}{\text{The probability of hypothesis H2}}$$

The *likelihood ratio* states the ratio between the conditional probabilities of E given H1, respectively, H2:

$$\frac{\text{The probability of evidence E given hypothesis H1}}{\text{The probability of evidence E given hypothesis H2}}$$

Finally, the *posterior odds* expresses the ratio between the conditional probabilities of the two hypotheses H1, respectively, H2 given evidence E:

$$\frac{\text{The probability of hypothesis H1 given evidence E}}{\text{The probability of hypothesis H2 given evidence E}}$$

Bayes's Theorem then says that the posterior odds equals the prior odds multiplied by the likelihood ratio:

$$\text{Posterior odds} = \text{Prior odds} \times \text{Likelihood ratio}$$

Below I initially assume that the two hypotheses H1 and H2 are not only incompatible but also jointly exhaustive. This assumption is far from innocent but for now it simplifies the explanation. Later in Section 6.1 I come back to it.

By itself Bayes's Theorem is just a mathematical equation. It derives its fame from its use as a way of thinking, as a way to update the probability of a hypothesis of interest after receiving new evidence. The probability of hypothesis H1 in the numerator of the prior odds is then called the *prior probability* of H1 and the

probability of H1 given E its *posterior probability*. The idea is that the prior probability of H1 is its probability *before* considering evidence E while its posterior probability is its probability *after* considering the evidence. In this way of thinking the likelihood ratio captures the *probative force* of evidence E in that it expresses how much more or less probable observing E makes hypothesis H1. It immediately follows from Bayes's Theorem that if the likelihood ratio of E with respect to H1 and H2 is greater than 1, then E makes H1 more probable than it was before receiving the evidence, while if the likelihood ratio is less than 1, then E makes H1 less probable than it was before receiving the evidence. When H1 is a guilt hypothesis, such as the hypothesis that Sally Clark killed her two sons, then evidence with likelihood ratio greater than, respectively, less than 1 can be called *incriminating*, respectively, *exculpatory*. Evidence with likelihood ratio 1 is *irrelevant*, since it does not change the prior probability of H: clearly multiplying the prior probability with 1 makes the posterior probability equal to the prior.

In applications to evidential reasoning we are always interested in the probability of a hypothesis of interest given the available evidence. Therefore, we are ultimately not interested in the posterior odds but in its numerator, that is, in the probability of hypothesis H1 given the evidence E. If we only have the value of the posterior odds but not the values of its numerator and denominator, then the numerator can be derived from the posterior odds as follows. Let H1 be a hypothesis the court is interested in, such as that John is the father of Mary's child. Then H2 is the hypothesis that John is not the father of Mary's child. We know that the conditional probabilities of a statement and its negation given the same condition add up to 1 (or to 100%). Let us also assume that after multiplying the prior odds with the likelihood ratio we arrive at a posterior odds of 3, so given the evidence E it is three times more probable that John is the father of Mary's child (H1) than that he is not (H2). This implies that the probability of H1 given E is 0.75 (or 75%), since 0.75/0.25 = 3. What we have done here is adding 1 to the likelihood ratio (1 + 3 = 4) and dividing the numerator by the result (3/4 = 0.75). This rule can always be used to deduce a so-called *posterior probability* from a posterior odds, provided the two hypotheses that appear in the odds are mutually exclusive and jointly exhaustive, as we are assuming for now.

This all looks very abstract, so let us illustrate it with concrete examples. Let us first consider the **blue and green taxis** example. Hypothesis H1 is that the taxi that hit the other car is blue while hypothesis H2 is that it is green. E is the witness testimony that he saw that the taxi was blue. The prior odds is then[3]

$$\frac{\text{The probability that the taxi is blue}}{\text{The probability that the taxi is green}} = \frac{0.15}{0.85} \approx 0.176$$

The likelihood ratio is

$$\frac{\text{The probability that the witness says the taxi is blue given that the taxi is blue}}{\text{The probability that the witness says the taxi is blue given that the taxi is green}}$$

$$= \frac{0.8}{0.2} = 4$$

[3] The symbol ≈ means 'is approximally equal to'.

So the witness testimony makes it four times more probable that the taxi is blue than before it was considered. We can now calculate the value of the posterior odds

$$\frac{\text{The probability that the taxi is blue given that the witness says the taxi is blue}}{\text{The probability that the taxi is green given that the witness says the taxi is blue}}$$

which equals 0.176 x 4 = 0.706. Then dividing 0.706 by 1.706 yields a posterior probability that the taxi is blue given the evidence of 0.414 or 41.4%, which is the probability we arrived at in the tabular method.

We see that in the odds version of Bayes's Theorem the base rate is expressed in the prior odds while the witness's reliability is captured in the likelihood ratio. Even though the witness testimony is incriminating since its probative force equals 4, the posterior probability that what the witness says is true remains less then 0.5 since the base rate of taxis is biased against what the witness says.

The drugs use and the paternity test examples can be analysed in the same way. I only show the calculation for the **paternity test example**. Assume again that other evidence has reduced the number of potential fathers to 10, of which John is one, so the prior probability that John is the father is 1/10 = 0.1. Then the prior probability that John is not the father is 0.9, so the prior odds is

$$\frac{\text{The probability that John is the father}}{\text{The probability that John is not the father}} = \frac{0.1}{0.9} \approx 0.11$$

The 99.99% = 0.999 test reliability yields the following likelihood ratio:

$$\frac{\text{The probability that John tests positive given that he is the father}}{\text{The probability that John tests positive given that he is not the father}}$$

$$= \frac{0.9999}{0.0001} = 9999$$

So the positive test makes it 9999 times more probable that John is the father. Multiplying this probative force with the prior odds yields 0.11 x 9999 ≈ 1100. Then 1100/1101 ≈ 0.999, which is a 99.9% probability that John is the father given the positive test.

In the two examples thus far the numerator and denominator of the likelihood ratio add up to 1. However, this does not have to be the case. Consider again the **tire tracks** example. With 5 million Dutch cars the prior odds that the accused's car made the tire tracks is

$$\frac{\text{The probability that the accused's car made the tracks}}{\text{The probability that another car made the tracks}}$$

$$= \frac{\text{1 in 5 million}}{\text{4,999,999 in 5 million}} \approx \text{1 in 5 million}$$

The likelihood ratio is

$$= \frac{\text{The probability of the match given that the accused's car made the tracks}}{\text{The probability of the match given that another car made the tracks}} = \frac{1}{1/5000} = 5000$$

Here the random-match probability is the denominator of the likelihood ratio. Its numerator equals 1 since if the accused's car made the tire tracks, its tire profile will certainly match with the tire tracks at the crime scene. We now compute the posterior odds as 5000 x 1 in 5 million, which equals 1/1000. Then the posterior probability that the accused's car made the tire tracks given the match is $1/1001 \approx 0.001$.

It is instructive to see what happens if the number of potential sources of the tire tracks can be reduced by other evidence. With 10,000 potential sources the prior odds is 1 in 10,000/9,999 in 10,000, which approximately equals 1 in 10,000. This yields a posterior odds of ½, which in turn yields a posterior probability of 1 divided by 1+2 = 1/3 or 33.3%. This can also be seen with the tabular method: with 10,000 potential sources and a random-match probability of 1 in 5000, we have that 2 cars will match with the tire tracks even though they are not the source. So the accused's car can be the source of the tracks or one of these other two cars, so the probability that it is the source is 1/3. If the number of potential sources is 5000, then the prior odds is approximately 1/5000, so the posterior odds is 1, so the posterior probability is 0.5 or 50%. Or in the tabular method we have that one of the 5000 cars will match without being the source while one is the source and the accused's car could be either of them. Finally, if the number of potential sources can be reduced to 500, then the posterior odds is 10, which yields a posterior probability of 10 divided by $10+1 = 10/11 \approx 91\%$. What this illustrates is that even if we have evidence with strong probative force, the posterior probability of a hypothesis given the evidence may range from low to high depending on the prior probability of the evidence.

This also explains what went wrong in the **child abuse** example. Now that we know the odds version of Bayes's Theorem, we see that the expert in fact reported the likelihood ratio of the observed injuries given two hypotheses. That he reported it in qualitative terms does not matter for the applicability of the theorem, since even without numbers we can recognise the court's mistake: it failed to show awareness that the likelihood ratio must be combined with the prior odds. As indicated above, the prior odds may be based on further evidence in the case (see also the discussion of the Denis Adams case below). It may well be that the additional evidence in the case justified a sufficiently high prior odds to warrant het court's conclusion but it is nevertheless worrying that the court showed no awareness that the prior odds is relevant, since this indicates that the court may have committed a fallacy.

I next return to the **Sally Clark** case. The unjustified independence assumption was not the only problem with the paediatrician's estimate that the probability that two children die from unexplained natural causes in a family such as the Clarks is 1 in 73 million. Many inferred from this very low probability that Sally almost certainly killed her two sons. I will now, following Dawid (2005, Section 4.3) show with Bayes's Theorem that this inference is based on an erroneous inversion of a conditional probability. Consider the hypotheses H1 that Sally Clark killed her two babies and H2 that she did not kill her two babies and the evidence E that both babies

died. Note that H2 does not imply that the babies died. Bayes's Theorem is thus instantiated as follows:

$$\frac{\text{The probability that Sally Clark killed her two babies given that they died}}{\text{The probability that Sally Clark did not kill her two babies given that they died}}$$

$$=$$

$$\frac{\text{The probability that Sally Clark killed her two babies}}{\text{The probability that Sally Clark did not kill her two babies}}$$

$$\text{x}$$

$$\frac{\text{The probability that the two babies died given that Sally Clark killed them}}{\text{The probability that the two babies died given that Sally Clark did not kill them}}$$

In determining the likelihood ratio the paediatrician's estimate can be used as the probability that the two babies died given that Sally Clark did not kill them. This yields a likelihood ratio of 73 million, since the babies will surely have died if Sally Clark killed them, so we have to divide 1 by 1 in 73 million, which equals 73 million. So the death of Sally Clarks two sons is strongly incriminating evidence, since after receiving this evidence Sally Clarks guilt is 73 million times more probable than before. However, the prior odds counters this strength, since the probability that Sally Clark killed her two babies is also very low: there are not many mothers who kill their children. On the basis of official murder statistics Dawid (2005) estimates this probability as 1 in 8.4 billion, admitting that his estimate is "at least as spurious" as the paediatrician's estimate. With Dawid's estimate the prior odds is 1 in 8.4 billion, which when multiplied with 73 million yields a posterior odds of 0.009. So the posterior probability that Sally Clark killed her two sons given that they died is negligibly small. And it becomes even 90 times smaller (0.0001) if the paediatrician's estimate is replaced by the later estimate of 1 in 850,000 by the experts who took the dependencies between the deaths into account.

One way to explain the fallacy committed by those who concluded from the paediatrician's estimate that Sally Clark almost certainly killed her sons is that they failed to see that we must compare the probability of *two* rare events: not only unexplained death of two babies by natural causes is rare but also a mother killing her two baby sons is rare. The rarity of the first event is accounted for in the likelihood ratio, which is high, while the rarity of double murder is expressed in the prior odds, which is low.

I finally apply Bayes's Theorem to the **Denis Adams** case, to illustrate the processing of multiple pieces of evidence and also to illustrate that probabilities cannot always be based on frequencies. Recall that in this case there were three pieces of evidence: the match between Denis Adam's DNA and DNA in the semen sample found in the rape victim (E1), the failure of the victim to recognise Adams in a lineup (E2) and the alibi provided by Adams' girlfriend that he had spent the night of the crime with her (E3). Let us consider as hypotheses that Adams was (H1), respectively was not the rapist (H2) (recall that we assume that the source of the semen was the rapist). How can Bayes's Theorem be applied to multiple pieces of evidence? If we can make an independence assumption then this is straightforward. The assumption is

that given the two hypotheses, the three pieces of evidence are statistically independent of each other. This is a generalisation of the definition of statistical independence given above in Section 3.3: we say that that statement A is statistically independent of statement B given hypothesis H if the conditional probability of A given H & B equals the conditional probability of A given H. In other words, for knowing whether A is true if H is true, it is irrelevant whether B is true or false.

Let us assume that the three pieces of evidence are statistically independent given H1 (and H2) in this sense. This assumption is far from obvious and needs to be argued; there no such thing as a general presumption that things are statistically independent unless shown otherwise; this is simply not how the world is in general. Nevertheless, the assumption allows us to illustrate how independent pieces of evidence can be processed through Bayes's Theorem. We are interested in the posterior probability that Adams was the rapist given the three pieces of evidence. We obtain the corresponding posterior odds by subsequently multiplying the prior odds with the likelihood ratios of each of the pieces of evidence:

$$\frac{\text{The probability that Adams was the rapist given E1 \& E2 \& E3}}{\text{The probability that Adams was not the rapist given E1 \& E2 \& E3}}$$

$$=$$

$$\frac{\text{The probability that Adams was the rapist}}{\text{The probability that Adams was not the rapist}}$$

$$\times$$

$$\frac{\text{The probability of E1 (the DNA match) given that Adams was the rapist}}{\text{The probability of E1 (the DNA match) given that Adams was not the rapist}}$$

$$\times$$

$$\frac{\text{The probability of E2 (the non-recognition) given that Adams was the rapist}}{\text{The probability of E2 (the non-recognition) given that Adams was not the rapist}}$$

$$\times$$

$$\frac{\text{The probability of E3 (the alibi) given that Adams was the rapist}}{\text{The probability of E3 (the alibi) given that Adams was not the rapist}}$$

In this way the posterior odds obtained after processing one piece of evidence functions as the prior odds for processing the next piece of evidence.

Having seen the analysis of the other examples, the reader will understand that a good way to estimate the prior is estimating the potential number of rapists. Should we include all adult males in Hemel Hempstead, or in the Greater London Area, or in England, the UK, Europe …? There is no easy answer to this question and the problem of determining the prior is often regarded as the Achilles heel of Bayesian thinking. Nevertheless, to make sense of the random-match probabilities estimated by the expert, we cannot escape the task to estimate the prior odds; as Bayes's Theorem show, without such an estimate nothing can be inferred from the match and its random-match probability about whether Adams was the rapist.

Let us, just to illustrate the reasoning, initially assume that there were 2 million potential rapists (perhaps all male adults in the greater London area). Then the prior odds is 1 in 2 million divided by 1,999,000 in 2 million, which, with a very small error margin that can safely be ignored, equals 1 in 2 million. For determining the likelihood ratio I go along with the defence's estimate of the random-match probability, which was 1 in 2 million. This yields a likelihood ratio of 2 million, since if Adams was the rapist, his DNA would certainly match. Clearly, multiplying 1 in 2 million by 2 million results in a prior odds of 1, so given only the DNA match it is just as probable that Adams was the rapist as that he was not the rapist: 50%.

So even with a very small random-match probability of 1 in 2 million it may, depending on the prior, not be probable that Adams was the source of the DNA, contrary to what is often concluded from a DNA match. Wrongly inverting a small random-match probability to conclude that the person who matches with a trace is almost certainly the source of the trace is sometimes called the *prosecutor's fallacy* (after Thompson & Schumann 1987), since several of the first instances of this fallacy in court were made by prosecutors presenting DNA evidence. In the paternity test and tire tracks examples we saw other instances of this fallacy.

To see the importance of the prior odds, let us now assume that there were 200,000 instead of 2 million potential rapists (perhaps all male adults in the Hemel Hempstead area). Then the probative force of 2 million of the DNA match must be multiplied with a prior odds of 200,000, which gives a posterior odds of 10, which in turns gives a posterior probability that Adams was the rapist given the match of 10 divided by 10+1 = 10/11 ≈ 91%. To some this may be sufficient to regard Adams guilty beyond reasonable doubt. However, it is important to be aware that even if after considering some evidence the posterior probability of guilt is very high, this does not mean that the case is closed, since new evidence could always bring the posterior down, even close to 0.

Let us illustrate this with processing the two other pieces of evidence. In doing so, I assume a prior of 1 in 200,000, so after processing the DNA match the posterior odds equals 10. We must now estimate the likelihood ratio of the non-recognition of Adams by the victim in a line-up (E2). This is not trivial either. Dawid (2005) estimates it as 1/9, so the non-recognition is nine times less probable if Adams was the rapist than if he was not the rapist. Dawid recognises that this estimate is speculative but let us go along with it. Then we obtain a new posterior odds of 10 x 1/9 = 10/9, which yields a new posterior probability that Adams was the rapist of approximately 53%. Next we must estimate the probative force of the alibi evidence. This is clearly weak since Adams' girlfriend had a reason to protect her boyfriend. Dawid tentatively estimates it as ½, so he regards it as the twice as probable that Adams girlfriend testified as she did if Adams was not the rapist than if he was the rapist. Then 10/9 multiplied by ½ yields 5/9 as the final posterior odds, which results in a final posterior probability that Adams was the rapist of 36%, clearly not enough to convict Adams.

The point of this analysis of the example is not to argue that Adams was wrongly convicted; for that conclusion the probability estimates are too speculative. What the example shows is that even though DNA evidence is strong incriminating evidence, it must always be combined not only with prior estimates but also with other evidence, some of which may be exculpatory. So a DNA match can never be a 'smoking gun' on its own.

There is more to say about this example. While random-match probabilities of DNA evidence are based on the frequency of occurrences of DNA profiles in a population, and while our estimates of the prior odds before processing any evidence could be based on frequency estimates of the number of male adults in the area, with the likelihood ratio's of the non-recognition and alibi evidence this was hardly possible. In these cases the probabilities are instead degrees of believability of statements about individual events or states of affair. In our example this means that after processing the non-recognition evidence, resulting in a new prior odds of 10/9 for processing the alibi evidence, we cannot say that we have reduced the number of potential sources of the DNA to, roughly, 2 persons. Instead, the probability of 53% that Adams was the rapist expresses a degree of believability of the claim that Adams was the rapist. A similar analysis is possible of the paternity test and tire tracks examples. For instance, in the paternity test example we may have evidence that more strongly points at John as the father than at the many other potential fathers, without being able to say that the number of potential fathers has been reduced. So if, say, we estimate a new prior probability of 10% that John is the father on the basis of other evidence, this does not necessarily mean that there are 10 potential fathers of which John is one; instead it means that we regard it 10 times more believable that someone else is the father than that John is the father. For such belief-type probabilities the mathematics is the same as for frequency-type probabilities, that is, Bayes's Theorem and the other laws of probability theory still apply.[4] However, the ways of justifying the probability estimates with which we calculate are different. I shall return to this issue in the next section.

6 Practical and theoretical issues with using probability theory in court

In this section I discuss several issues concerning the use of probability theory in court. First, above I made two assumptions that are not always satisfied in practice, namely, that the hypotheses we compare with Bayes's Theorem are not only mutually exclusive but also jointly exhaustive and that multiple pieces of evidence are statistically independent of each other given the hypotheses that are compared. Another issue is that there is a danger that the mathematical form in which statistical evidence is presented creates an unfounded impression of objectivity. I now discuss these three issues in turn.

6.1 Non-exhaustive hypotheses

If the hypotheses H1 and H2 that are compared in Bayes's Theorem are not exhaustive, then we cannot derive a posterior probability from a posterior odds. The reason is that in that case the two probabilities of H1 given evidence E and H2 given evidence E do not add up to 1 (or to 100%), since another hypothesis may be the true

[4] This point is in fact not entirely uncontested; there are non-standard schools of thought that claim that the mathematics of belief-type probabilities is different than for frequency-type probabilities. See e.g. Section 2 of the introduction of Prakken et al. (2020) and the references therein.

one. This is a serious problem, since in court we are in the end always interested in the probability of a hypothesis of interest given all available evidence. At first sight, there would seem to be a simple way to avoid this problem by always letting H2 be the negation of H1. This is what we did in most of our examples. However, there is a practical problem with this solution, since in many cases it is hard to estimate the probability of a piece of evidence given the negation of a hypothesis (which we must do to determine the likelihood ratio). For example, suppose some morning we see that our car, which is parked along the street in front of the neighbours' home, is damaged, and we want to consider the hypothesis that another car hit it. Based on our commonsense we can make a reasonable estimate of the probability of the damage if another car hit our car, since that is a good explanation of the damage. However, doing the same for the hypothesis that no other car hit our car is much more difficult, since by itself this negative fact does not give a good explanation of the damage. So it is tempting to compare our first scenario with a specific other scenario, such as that a heavy object fell off a truck while it passed our car. However, this creates the danger that other possible scenarios are overlooked, so that the step from posterior odds to posterior probability cannot be made. In our example yet another explanation might be that a sailboat mast which the neighbours had stored at the top of their flat roof fell off the roof in a storm (this actually happened to my car a few years ago).

Another example of this danger is a Dutch criminal case[5] in which a man was shot and killed in a home and DNA that matched with DNA of another man who was known to be in the home at the time of the killing was found on the victim's body. In this case it was beyond dispute that the DNA found on the victim was the other man's DNA and the issue was how the other man's DNA had ended on the victim's body. The prosecution's hypothesis was that this had happened by direct violent contact while the defence claimed that it had happened by secondary transfer in the home, for instance, since the victim had touched an object that was previously touched by the other man. Unlike the issue who is the source of a DNA trace, in which the random-match probability can be based on statistics, the issue of how the other man's DNA had ended up on the victim's body cannot be analysed in terms of statistics. Accordingly, a DNA transfer expert estimated a likelihood ratio in qualitative terms, testifying that he regarded it as much more probable that the DNA was on the victim's body because of direct violent contact than because of secondary transfer in the home. The court, perhaps trained in Bayesian thinking, refused to draw a conclusion from this testimony on the grounds that the expert should also have considered another explanation for the DNA trace. What had happened in the case is that after the victim was killed, he was wrapped in a tapestry and transported to a place in the countryside in the other man's car: according to the court the other man's DNA might have been transferred to the dead victim's body during this car ride.

[5] ECLI:NL:GHARL:2014:8932

6.2 Non-independent evidence

Our application of Bayes's Theorem to the Denis Adams case showed that if multiple pieces of evidence are statistically independent of each other given the hypotheses, then processing them is straightforward: just estimate their individual likelihood ratios and then multiply them with each other and the prior odds. However, in non-trivial cases evidence is often non-independent of each other. We already saw an example in the Sally Clark case, where the two deaths of he babies could have been due to underlying shared genetic, domestic or social causes. To give just a few other examples, imagine a case where there is camera footage of a man looking like the suspect in the main hall of Utrecht Central Station at 10pm and a witness testifies that she saw a man looking like the suspect leaving the station at 10.15pm on the same day. Clearly, the probability that a witness will observe a man looking like the suspect given that the man is the suspect is lower than the probability that the witness will observe a man looking like the suspect given that the man is the suspect and a man looking like the suspect is on camera footage of 10 minutes earlier. So the likelihood ratios of these pieces of evidence cannot be multiplied. Or suppose that DNA matching a suspect's profile is found on the victim's shirt and on furniture in the room where the suspect was killed. Then the probability of a match given that the suspect is not the source is lower than the probability of a match given that the suspect is not the source but DNA with the same profile was found at the furniture: in the later case it is probable that a person with the same DNA profile as the suspect was in the room, so finding another trace of the same DNA becomes more probable, whether the DNA is of the suspect or of another person with the same profile.

The bad news in such cases is that then application of Bayes's Theorem becomes much more complicated. It follows from Bayes's Theorem that we must then for a given piece of evidence E estimate the following ratio:

$$\frac{\text{The probability of E given hypothesis H1 } \textit{and all other evidence on which E depends}}{\text{The probability of E given hypothesis H2 } \textit{and all other evidence on which E depends}}$$

Clearly, determining these probabilities is often much more difficult than determining the probabilities of E given H1 or given H2 alone. It should be said that there are more sophisticated ways to process probabilities, which use graph theory to graphically represent statistical independence relations. These so-called Bayesian networks are currently very popular in artificial intelligence, for instance, for medical applications. However, Bayesian networks are for non-specialists much harder to understand than simple uses of Bayes's Theorem. This is a serious practical obstacle to their use in legal contexts (although some claim that this problem is not unsurmountable; see e.g. Fenton & Neil 2011).

6.3 Justifying probability estimates

So far we have mainly focused on how probabilities reported by experts in court cases can be used to draw useful conclusions. However, when courts are confronted with such reports, they also face the question to which extent the probabilities reported by the experts are justified. This is, of course, a special case of what courts must always do when confronted with expert evidence, but a special feature of statistical evidence is that it is often presented in mathematical form, which may create an unfounded

impression of objectivity. Non-specialists are often not aware of the fact that mathematical formulas are nothing else than statements in some language and that they can be false just as any statement in natural language can be false. When calculating with probability estimates it holds that 'garbage in, garbage out': if the probabilities that go into a calculation cannot be justified, then the probabilities derived from them will not be justified either, even if the way they were derived from the 'input' probabilities is mathematically sound.

So what can justify probability estimates? One way is to base them on *statistical frequencies*, as is usually the case with random-match probabilities of forensic trace evidence. They may also be based on *scientific experiments*. For example, controlled experiments have been done on the probative value of recognitions in police lineups. When probabilities cannot be based on statistics or experiments, then *professional expertise* is a good alternative, such as in the child abuse example, or in the DNA transfer case discussed in Section 6.1. Of course, appeal to expert opinion is defeasible, since experts can be mistaken, multiple experts can disagree, experts may be tempted to make statements that are outside their expertise, or their claim that they are experts may be exaggerated or even false. Nevertheless, professional expertise is often a reasonably reliable source of statistical evidence.

However, expertise is not always available. This often holds for prior probabilities, since experts usually only have expertise about the probative force of certain types of evidence and therefore usually withhold judgement on the prior. It also holds for probabilities concerning non-technical or non-medical matters from daily life, which often arise in legal cases. For instance, suppose that in the paternity test case there is evidence that John and Mary dated each other a few times. *Commonsense* tells us that this is incriminating evidence but it is very hard to express this in reliable numbers. Moreover, commonsense is not a very reliable knowledge source and can shift into subjective opinion or even prejudice.

6 Conclusion: the benefits and limitations of using probability theory in court

Summarising, we have seen that the main benefits of probability theory, especially of the Bayesian way of using it, are pedagogical and therapeutic. It is easy to show with simple examples that various seductive forms of reasoning with probabilities are fallacious. Therefore, basic knowledge and understanding of (standard and Bayesian) probability theory is important for legal professionals, scholars and students to help recognising and avoiding fallacies when interpreting statistical evidence. For more on such uses of Bayesian thinking see e.g. Fenton & Berger (2016) and Dahlman (2020). However, Bayesian probability theory is less suitable as a general way of thinking about legal evidential issues. The Bayesian way of thinking is for many people counterintuitive, precise probability estimates are often hard to give and statistical non-independence issues seriously complicate matters in non-trivial cases.

This raises the issue of how fact finders should embed the statistical evidence presented by experts into their general thinking about a case if their general thinking cannot be in terms of probability theory. I believe that much would be gained if

courts[6] in their decisions show an awareness when referring to likelihood ratios or random-match probabilities that these must be combined with an assessment of the prior odds on the basis of other evidence, where this assessment can very well be in non-numerical terms and obtained with a non-probabilistic way of thinking. For example, in the child abuse case the court could have said:

> 'The court accepts the conclusions of expert (…) that the combination of the brain injury, the bruises under the hard meninges, the retinal haemorrhages and the skin lesions is very much more probable in a non-accidental event than in an accidental event or medical condition. *Moreover, on the basis of the other evidence the court regards a non-accidental event as not much less probable than an accidental event or medical condition.* The court infers from this that the victim must have been seriously assaulted'.

With the italicised phrase the court would have shown awareness that the expert's likelihood ratio must be combined with a prior odds in order to draw conclusions about the hypotheses it is considering. The court could then explain in non-probabilistic terms why it regards a non-accidental cause of the injuries as not much less probable then the alternative hypothesis.

An interesting question is what are rational constraints on such a non-probabilistic way of evidential reasoning. The two main alternatives that have been proposed in the academic literature are argumentation-based and scenario-based thinking. Argumentation-based approaches take arguments, more specifically, series of inferences from evidence to conclusions, as the main concept. This approach goes back to Wigmore's (1931) charting method for legal proof and was revived in the 1980s and 1990s by the so-called New Evidence scholars (cf. Anderson et al. 2005). The idea of this method is that making the various inferences in an argument explicit allows one to identify sources of doubt in these arguments.

While the latter is a strong point of argumentative thinking, a problem is that it does not explicitly allow for the construction and comparison of alternative scenarios as a way of maintaining overview of a mass of evidence. This is a strong point of scenario-based (sometimes also called story- or narrative-based) thinking, which consists in constructing and comparing multiple plausible and coherent scenarios that explain the evidence (see e.g. Van Koppen & Mackor 2021). The scenario that is the most plausible and coherent and that best explains the evidence should be accepted as true. Scenario-based thinking is thus a form of what philosophers have called inference to-the-best explanation. Some, though not all scholars see inference-to-the best explanation as a qualitative approximation of Bayesian thinking, where judgements on how well the scenarios explain the evidence are the counterpart of likelihood ratios and considerations on the scenarios' plausibility and coherence are the counterparts of the prior odds (cf. Jellema 2019). If this is true, then scenario-based thinking could be a good overall way for fact finders to structure their thinking while they can zoom in on specific issues with argumentation or probability theory when appropriate. However, the academic debate on this issue is still ongoing, witness e.g. the papers in Prakken et al. (2020).

[6] This does, of course, not apply to juries, which generally do not have to give reasons for their judgements.

References

Anderson, T.J., Schum, D.A. & Twining, W.L. (2005). *Analysis of Evidence*, second edition. Cambridge: Cambridge University Press.

Dawid, Ph. (2005). Probability and proof. Online appendix to "Analysis of Evidence" by T.J. Anderson, D.A. Schum and W.L. Twining. http://www.statslab.cam.ac.uk/~apd/

Dahlman, C. (2020). De-biasing legal fact-finders with Bayesian thinking. *Topics in Cognitive Science*, 12(4), 1115-1131.

Derksen, T. & Meijsing, M. (2009). The fabrication of facts: the lure of the incredible coincidence. In H. Kaptein, H. Prakken & B. Verheij, (eds.): *Legal Evidence and Proof: Statistics, Stories, Logic (Applied Legal Philosophy Series).* Farnham: Ashgate, 2009, pp. 39-70.

Fenton, N. & Berger, D. (2016) Bayes and the law. *Annual Review of Statistics and Its Application*, 3, 51–77.

Fenton, N.E. & Neil, M. (2011). Avoiding legal fallacies in practice using Bayesian networks', *Australian Journal of Legal Philosophy* 36, 114-151, 2011.

Jellema, H. (2019). Case comment: responding to the implausible, incredible and highly improbable stories defendants tell: a Bayesian interpretation of the Venray murder ruling. *Law, Probability and Risk* 18: 201-211.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin, London.

Koppen, P. J., Van & Mackor, A. R. (2021). The scenario theory about evidence in criminal law. To appear in: Dahlman, C. Stein, A., Tuzet, G. (eds.): *Philosophical Foundations of Evidence Law*. Oxford: Oxford University Press.

Lempert, R. (1986) The new evidence scholarship: analyzing the process of proof. *Boston University Law Review* 1986-66: p. 439-477.

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

Prakken, H., F.J. Bex & A.R. Mackor (eds.) (2020), *Topics in Cognitive Science* 12:4, Special issue on *Models of Rational Proof in Criminal Law*.

W.C. Thompson & E.L. Schumann (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.

Tribe, L. (1971), Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, 84: 1329–1393.

Tversky, A. & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science* 185: 1124–1131.

Wigmore, J.H. (1931) *The Principles of Judicial Proof*. Boston: Little, Brown and Company, 2nd edn.