

Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio

Franco Taroni, Silvia Bozza, Alex Biedermann, and Colin Aitken

University of Lausanne, Ca' Foscari University of Venice and University of Edinburgh

Abstract. The use of the Bayes factor or likelihood ratio as a metric to assess the probative value of forensic traces is largely supported by operational standards and recommendations in different forensic disciplines. However, the progress towards more widespread consensus about foundational principles is still fragile as it raises new problems about which views differ. It is not uncommon, for example, to encounter scientists who feel the need to compute the probability distribution of a given expression of evidential value (*i.e.*, a Bayes factor), or to place intervals or significance probabilities on such a quantity. The paper here presents arguments to show that such views involve a misconception of principles and abuse of language. The conclusion of the discussion is that, in a given case at hand, forensic scientists ought to offer to a Court of justice a given single value for the Bayes factor, rather than an expression based on a distribution over a range of values.

Key words and phrases: Bayes factor, Likelihood ratio, Probability, Estimation, Uncertainty, Variability.

1. INTRODUCTION

Decades ago, discussions about how to assess the probative value of forensic traces, such as DNA or other transfer material, were less structured and formalized than they are today, and the diversity of opinion was more substantial. Today, there still are instances of disagreement, which leads some commentators to endorse spurious methodological pluralism (*e.g.*, Simmross (2014)), but on the whole it appears reasonable to say that the number of scientists or lawyers who still are skeptical about the metric called Bayes factor, in the context often referred to as the likelihood ratio, has reduced considerably. This development is favoured and supported by the fact that operational standards and recommendations in different forensic disciplines substantially support the use of this metric of probative value. This progress towards more widespread consensus about foundational principles is encouraging, but it is still fragile as it raises new problems about which views differ.

University of Lausanne and Ca' Foscari University of Venice (e-mail: franco.taroni@unil.ch; silvia.bozza@unive.it). University of Lausanne and University of Edinburgh (e-mail: alex.biedermann@unil.ch; C.G.G.Aitken@ed.ac.uk).

In this paper, two such problems are presented and discussed. One concerns the understanding of the notion of probability as a measure of uncertainty, the other concerns the use in evidence evaluation of the Bayes factor which in some cases is represented by a ratio among probability density functions. An issue that is common to these two problems is the idea held among some forensic scientists (*e.g.*, [Stoel and Sjerps \(2012\)](#), [Alberink and Bolck \(2008\)](#)) that both probability and Bayes factors can, in some way, be ‘estimated’ and, thus, be accompanied with some kind of ‘interval estimate’ of a true value. The choice of this topic for discussion is motivated by the fact that it is not uncommon to encounter forensic scientists who argue the need to determine the probability distribution of a given expression of evidential value (*i.e.*, a Bayes factor), or to fit an interval to such an expression.

The paper argues that such views may involve a misconception of principles and an abuse of language. It will be shown that with a clear view on what the concepts of probability and of the Bayes factor, really mean and do *not* mean, the widespread use of terms such as ‘estimation’ to justify the use of intervals surrounding given values turns out – in many situations – to be internally inconsistent, and hence misconceived.

The objective of this paper is to clarify the definitions of some basic terms and their applications in forensic science in relation to the topic of value of evidence. In particular, distinctions will be drawn between the concepts of uncertainty and variability and the concepts of estimation and assignment. The conclusion of the discussion is that, for a particular case, forensic scientists ought to offer to the Court a single value for the Bayes factor, rather than a range of values.

The structure of this paper is as follows. In Section 2 the notion of the evaluation of scientific findings is introduced, followed by the definition of the Bayes factor and its use in forensic science. Section 3 presents an overview of the subjective interpretation of probability, while issues related to the practical derivation of the Bayes factor are considered in Section 4. Particular attention will be paid to what, in informal discussion, is often vaguely referred to as ‘uncertainty’ about the actual value of a Bayes factor. Section 5 will pursue this discussion with the use of practical examples, emphasising that there is a fundamental difference between the concepts of uncertainty and variability and that this difference should be kept in mind in discourses about the likelihood ratio. Conclusions are presented in Section 6.

2. THE PROBLEM: THE EVALUATION OF SCIENTIFIC FINDINGS

The assessment of the value of scientific findings is commonly considered through the derivation of a Bayes factor, a rigorous concept that provides a balanced measure of the degree to which the evidence is capable of discriminating among competing propositions suggested by opposing parties at trial. Denote by H_p and H_d the propositions of interest to the Court, which need to be mutually exclusive but need not be exhaustive, where the subscript p stands for the prosecution’s proposition and the subscript d stands for the defence proposition. The scientific outcome, denoted $E = \{y, x\}$, is given by the observations or measurements y on questioned material (*e.g.*, traces or trace material recovered on a crime scene) of unknown origin, often known as *recovered material*, and the observations or measurements x on *control material* (*e.g.*, material belonging to

the suspect) of known origin. The odds form of Bayes theorem enables the prior odds¹ in favour of the prosecution's proposition, $Pr(H_p)/Pr(H_d)$, to be updated to the posterior odds given the findings E , $Pr(H_p | E)/Pr(H_d | E)$, that is

$$(2.1) \quad \frac{Pr(H_p | E)}{Pr(H_d | E)} = \frac{L(H_p; E)}{L(H_d; E)} \times \frac{Pr(H_p)}{Pr(H_d)},$$

omitting explicit mention of background information I from the probability and likelihood statements for shortness of notation (and throughout the remainder of the paper). The fraction $L(H_p; E)/L(H_d; E)$ is written as the ratio of the likelihood $L(H_p; E)$ of H_p given E to the likelihood $L(H_d; E)$ of H_d given E , often represented as $Pr(E | H_p)/Pr(E | H_d)$. The ratio of the posterior odds and the prior odds is the Bayes factor (BF). It measures the change produced by a given item of information in the odds in favor of one proposition as opposed to another, when going from the prior to the posterior distribution. Determination of the Bayes factor is typically considered to be in the domain of the forensic scientist and its value implies either an increase or decrease in the prior odds once forensic findings are taken into account. In forensic science applications, this quantity is usually referred to as the likelihood ratio, though it is emphasised that the Bayes factor does not always simplify to a ratio of likelihoods². When the competing propositions of interest H_p and H_d are 'simple', that is there is only one main proposition and one alternative, the Bayes factor reduces to the well-known likelihood ratio (LR), and depends only on the observed data. When 'composite' propositions are considered (e.g., when at least one of the compared propositions covers several possibilities), the Bayes factor does not reduce to the likelihood ratio. Thus, generally speaking and as a first example, evaluation of a simple main proposition versus a simple alternative proposition is only a special case. Often, scientists may face the more general situation of dealing with at least one composite hypothesis. In the context of forensic science, this occurs in scenarios that involve multiple hypotheses for at least one of the party's propositions. With DNA stains, for example, the range of potential donors other than a particular suspect at hand may not only cover unrelated individuals, but also persons who may share various degrees of relatedness with the suspect. As a second example, consider the case where the competing propositions embrace composite hypotheses about parameters of interest (e.g., the question of whether the proportion of items that contain an illicit substance is greater or lower than a given threshold). In such a case, the Bayes factor can be expressed as a ratio of weighted likelihoods (Kass and Raftery, 1995). It no longer depends only upon the sample data. In view of this, the term likelihood ratio will be used as a synonym for Bayes factor for the rest of the paper, to include the wider use of the first example in forensic science applications.

Notice that to compute the likelihood ratio in a particular application, such as the evaluation of findings in forensic science, two ingredients are necessary: the probability distribution of the outcomes if hypothesis H_p is true, and the

¹ The term 'odds' is used loosely here as the propositions need not be exhaustive.

² Note that the term likelihood must not be treated as a synonym of probability (Lindley, 2006). The intuitive reason for the term likelihood is that it measures the ability of an explanation to give rise to the collected findings; available measurements for which the likelihood is large are more likely to occur under a given explanation rather than under an explanation for which the likelihood is small.

probability distribution of those outcomes if hypothesis H_d is true. This procedure gives rise to a series of questions. Some discussants ask, for example: Does there exist a *true value* of the likelihood ratio that could in some way be estimated? Is it possible to compute a credible interval or a confidence interval on it?

In reply to these questions it is first necessary to look more closely at the definition of the likelihood ratio. In practice, a forensic scientist may deal with rather different contexts and the likelihoods at the numerator and denominator that need to be ‘computed’ may have a different interpretation, and therefore so may the likelihood ratio itself. Consider the case where a biological stain has been found at the crime scene and that DNA analyses performed by a forensic laboratory have led it to report a match E between the genetic profiles characterizing the recovered material, y , and the control material x found on a suspect, respectively. The competing propositions of interest to the Court may be:

H_p : the suspect is the origin of the trace;

H_d : another person, unrelated to the suspect, is the origin of the trace.

The likelihood ratio is therefore given by the ratio between the *probability* of a reported match E if the hypothesis H_p is true, versus the probability of a reported match E if the hypothesis H_d is true, that is $LR = \Pr(E | H_p) / \Pr(E | H_d)$.

Usually, the numerator is not the main issue of discussion. With some reasonable assumptions it can often be taken to be equal to 1. Typically, if the hypothesis H_p is true, then it is generally assumed the laboratory analyses will report a match. This simplifies the likelihood ratio to $1 / \Pr(E | H_d)$. The conditional probability of the denominator, usually denoted by γ for short, is clearly not available at a glance, so the problem of the expert is to assess this probability. But, how should this probability be interpreted? Does it make sense for a person to speak in terms of estimation (either point or interval estimation) of the evidential value, which is in the form of a likelihood ratio which is the expression of a ratio of two probability density functions? Related to this question is consideration of a (prior) probability distribution for the probability³ in the denominator, in order to enable a subsequent Bayesian learning procedure. Would such consideration permit the user to advocate an ‘estimation’ of the likelihood ratio? Moreover, would this imply the specification of a probability distribution for the likelihood ratio? These are a series of closely intertwined questions, and their answers are crucial to clarify how the likelihood ratio may be interpreted, and how claims of ‘uncertainty’ about its value should be handled.

Clearly, it is not always possible to express the likelihood in terms of probabilities. Forensic scientists often deal with cases where the measurements on questioned items are on a continuous scale (e.g., the refractive index of glass fragments). Whenever data are continuous, a probability density function $f(\cdot)$ must be specified since it is not possible to assign probabilities to specific values of a continuous random variable. Consider the case where a window is broken during the commission of a crime, and measurements on the refractive index of glass fragments recovered on a suspect’s pullover (y), are compared to those of the broken window’s glass fragments (x). The competing propositions of interest to the Court may be:

³ Note the implied phrase ‘probability of a probability’, which is unsound from a philosophical point of view (Lindley, 2000)

H_p : the recovered and the control glass fragments originate from the same source;

H_d : the recovered and the control glass fragments originate from different sources.

Continuous measurements are being considered and the likelihood ratio may be written as

$$(2.2) \quad LR = \frac{f(x, y | H_p)}{f(x, y | H_d)},$$

where probabilities have been replaced by densities. Note that observations do not always conform to standard models (e.g., a Normal distribution), and statistical methods can be implemented to estimate the unknown density function $f(\cdot)$, but does this imply an estimation of the likelihood ratio?

All these fundamental questions give rise to three points of discussion that will be addressed in the succeeding sections. These points are (a) the interpretation of probability (Section 3), (b) the notion of estimation and its role in relation to the likelihood ratio (Section 4), and (c) the conceptual difference between *uncertainty* and *variability* illustrated by the derivation of a likelihood ratio distribution (Section 5). The conclusion (Section 6) summarises the ideas discussed in the context of the two main problems of the understanding of probability and the communication of a likelihood ratio.

3. INTERPRETATION OF PROBABILITY

As outlined in the previous section, the likelihood ratio is often expressed as a ratio between two conditional probabilities: the probability of observing given findings if the proposition H_p is true, and the probability of observing the same findings if the hypothesis H_d is true. Therefore, the task of the expert is focussed on the assessment of these probabilities. How to think sensibly about probability? This may seem an easy or almost unnecessary question as most scientists will have at least some idea in their minds when they think of the term ‘probability’. A few comments may be worthwhile in this context because it is important to distinguish between, on the one hand, the definition of probability and, on the other hand, the interpretation of the abstract (mathematical) concept of probability (e.g., [Hays and Winkler, 1970](#)). It is well known to scientists that, axiomatically, probabilities take values in the range between 0 and 1, including these limiting values, and independently of the way in which probabilities are actually interpreted. The value 0 corresponds to an event whose occurrence is impossible, and probability 1 is reserved for an event whose occurrence is certain. While that is not controversial, the main question of interest is ([Aven and Reniers, 2013](#)): what does it mean for somebody to specify a probability p for a given event? What does such a value *really* mean?

To answer this question in both a meaningful and operationally sound way, it is perhaps useful to take one step back and ask why one uses probability. One immediate answer, which is widely supported, says that probability serves as measure of uncertainty for something – an event or proposition – that is unknown to a given reasoning individual. For the individual at hand, it is not known whether a given proposition (for example, the suspect is at the origin of the trace), or a given event (for example, a random person from a relevant population has a given DNA sequence) is true or not. Given the partial knowledge of the

individual, the value p expresses the degree to which the proposition or event of interest is taken to be true. Probability provides a very close description of the problem faced by any person contemplating an uncertain event or proposition.

Two immediate consequences follow from this view. First, probability depends on the extent of one's knowledge, also sometimes referred to as a knowledge base. A given event or proposition may be known to be true for one individual (who, arguably, would give this proposition a value $p = 1$), but not so for another individual. The latter would thus have a probability smaller than 1, $0 < p < 1$. Arguably, probability is *conditional* on available information and there is no problem in principle with two persons having different probabilities for the same event (Lindley, 2000). This emphasis on conditionality leads to a second important point, namely that probability is personal. It is personal in the sense that it refers to the subject who holds a particular probability, hence the alternative designation 'subjective probability'. Here, the term 'subjective' refers to the opinion of the person of interest, and not to a suggestion of arbitrariness.

As an intermediate summary, note that probabilities reflect the extent to which an individual's knowledge is imperfect. It is important to acknowledge that such personal belief is graduated. One can believe in the truth of an event more or less than in the truth of another one, at any given time, essentially because findings can accumulate and modify one's probability of the event's happening. Probabilities depend necessarily on the individual, on available information, may change as the information changes, and may vary amongst individuals because different individuals may have different information or assessment criteria.

A corollary of these considerations is that probabilities are not states of nature, but states of mind associated with individuals. One definition of probability is to consider it as a *measure of belief* in the outcome of an event or the truth of a proposition. The measures of belief held by an individual need to conform to the rules of probability to be coherent. Coherence has the normative role of forcing people to be honest and to make the best assessments of their own measures of belief. Note, however, that the axiomatic foundation of probability theory only places constraints on the values that probability may take (*i.e.*, a value between 0 and 1), and how probabilities ought to be combined (*i.e.*, with the rules of addition and multiplication), but it does not say anything about what ought to be one's probability in a particular instance.

Acknowledgement that probabilities are states of mind raises the issue of how to elicit personal probabilities. There are various devices and procedures to do so. For the purpose of the current discussion, it will be sufficient to illustrate one example. It focuses on the measure of probabilities as exhibited by an individual in terms of the bets that the individual is willing to accept. According to this scheme, an individual's probability for a proposition is elicited by comparing two lotteries of the same price. For example, define a first Lottery A as winning 100 (units of some sort of value, which may be monetary) with probability $p = 0.5$, and winning 0 with probability $(1 - p) = 0.5$, and a second Lottery B as winning 100 if it rains tomorrow, and winning 0 if it does not rain tomorrow. Here, 'rain tomorrow' is the proposition about which you are uncertain. Note that it could be any other event, including events of the present or the past. The fact that the scheme is based on a gamble should not be objectionable because, at worst, the gain will be zero in either gamble, so the individual need not fear a loss and

hence refuse to engage in the gamble. With this condition, one can assume that one would choose that offer which presents the greater chance of winning the prize. Clearly, if one prefers Lottery B, then this signifies that one considers the probability of rain tomorrow to be greater than 0.5. By analogy, choosing Lottery A would imply that one's belief in rain tomorrow is lower than 0.5. Moreover, in a case in which one is indifferent between the two gambles, one's probability for rain tomorrow would equate to the probability of winning the prize in Lottery A. One can then conceive of a procedure in which one adjusts the chance of winning in Lottery A so that the individual, whose probability for a proposition of interest is to be elicited, would be indifferent with respect to Lottery B. Also, one can proceed similarly to elicit an individual's personal probability for any other event of interest.

According to this scheme, typically found in Lindley's writings, one's probability is thus defined by *a point of indifference* with respect to a standard. Clearly, one can adjust the measure of belief of success in the reference gamble in such a way that one will be indifferent with respect to the truth of the event about which one needs to give one's probability. This understanding is fundamental, as it implies that *probability is given by a single number*. It may be hard to define, but that does not mean that probability does not exist in an individual's mind (Jeffreys, 1931). One cannot logically have two different numbers because they would reflect different measures of belief.

This interpretation of probability is more typically known as the *belief type* interpretation of probability. It is a very common and, for various reasons, appropriate interpretation of the theory (de Finetti, 1974). It considers all probabilities as subjective, in the sense of personal expressions of degree of belief held by an individual. There are other interpretations of probability, though they run rather quickly into conceptual difficulties and may exhibit drawbacks in practice, in particular in forensic and legal contexts (Lindley, 2006).

According to the *classical* definition of probability (Laplace, 1814), for example, the probability of an event about which we are uncertain and that can result in several mutually exclusive and equally likely outcomes is given by the fraction of the number of outcomes that are favorable to it⁴. This definition is simple and intuitive, but circular since it contains the term 'likely' as part of the definition, and clearly deficient in the forensic context of interest since for the cases under attention such a split into equally uncertain possibilities does not exist (Lindley, 2006).

There is also a *frequency-based* interpretation of probability, where probability is interpreted using the concept of frequency rather than beliefs.

As an illustration of the connection between frequency and probability, consider an urn containing a certain number of balls, indistinguishable except by their colour, which is either white or black, and the number of balls of each colour being known. The extraction of a ball from this urn defines an experiment having two and only two possible outcomes that are generally denoted as success (say, the withdrawal of a white ball) or failure (say, the withdrawal of a black ball). Let B denote the event 'a white ball is extracted'. Under the circumstances that balls are all indistinguishable from each other except for the colour, the

⁴The concept of equally likely, at the basis of the classical definition of probability, is discussed in Hacking (1975)

subjective probability to extract a white ball can be assessed as the known *proportion* θ of white balls, that is $Pr(B | \theta) = \theta$. Assuming the urn contains a large number of balls, so that the extraction of a few balls does not alter its composition substantially, individual draws (*i.e.*, sampling⁵) will be considered as with replacement and the probability of extracting a white ball at subsequent withdrawals will still be θ , independently on previous observations. In this way one realizes a series of Bernoulli trials, where the outcome of each trial has a constant probability independent from previous outcomes. Suppose now the observer does not know the absolute value of balls present, nor the proportion that are of each colour. De Finetti showed that every series of experiments having two and only two possible outcomes that can be taken as exchangeable, that is the probability assigned to the outcomes of a sequence of trials is invariant to permutation, can be represented as random withdrawals from an urn of unknown composition. If one can assess one's uncertainty in such a way that labelling of the trials is not relevant, then it can be proved that as the number of observations increases the relative frequencies of successes (*i.e.*, the relative frequency of white balls) tend to a limiting value, that is the proportion θ of white balls. A *subjective* assessment about the outcome of a sequence of Bernoulli trials is equivalent to placing a prior distribution on θ . According to this, one only needs to model a prior distribution $Pr(\theta)$ for the various possible values of θ : personal beliefs concerning the colour of the next ball extracted can be computed as

$$(3.1) \quad \begin{aligned} Pr(B) &= \int_{\theta} Pr(B | \theta)Pr(\theta)d\theta \\ &= \int_{\theta} \theta Pr(\theta)d\theta \end{aligned}$$

The introduction of a prior probability distribution modeling personal belief about θ may seem, at first sight, in contradiction with the previous statements that the probability is a single number. One can, in fact, have probabilities for events, or probabilities for propositions, but not probabilities of probabilities, otherwise one would have an infinite regression (de Finetti, 1976). Confusion may arise from the fact that parameter θ is generally termed as '*probability of success*'. However, it is worth noting that, although it is effectively a probability, it represents a *chance* rather than a *belief*.

To understand how to pass from observations of individual trials (which may be summarised in terms of the relative frequency) to a probability distribution over θ , it will be helpful to describe how a person can revise, as new information becomes available, her prior beliefs about the possible occurrence of the binary trial discussed so far. Suppose that the initial state of mind of the individual may be described by an urn containing w white balls and b black balls: the individual's uncertainty about the outcome of the experiment can be described as the probability to observe a ball of colour white after one draw from this urn and may be quantified in $w/(w + b)$. A simple method to revise personal probabilities is represented by Polya's learning scheme, according to which after

⁵Note that the use of the term 'sample' in this context is one of purely technical nature in statistics and has nothing to do with the widespread but inappropriate use of the same term for designating physical trace material recovered or collected in forensic science context. In particular, seizing (*e.g.*, at crime scenes) and analysing trace material has to deal with the material as it is, irrespective of its condition; there is no such thing as randomisation, for example.

each drawing, the extracted ball is returned together with a ball of the same colour (i.e., sampling is with replacement and duplication). In this way, the observation of a ball of a given colour increases the degree of belief of observing a ball of the same colour at the following extraction. According to this learning scheme the probability to observe a white ball at the second extraction will be conditional on the outcome of the previous one. It will be quantified as $(w + 1)/(w + b + 1)$ whenever a white ball was observed at the first draw; conversely, if a black ball was observed, it will be quantified as $w/(w + b + 1)$. In the same way, one can compute the probability of any sequence of outcomes for any number of draws. The subjective learning that is realized following this scheme is the same as if one's belief about θ , described by a probability distribution $Pr(\theta)$, is modelled by a beta distribution with parameters $\alpha = w$ and $\beta = b$. In fact, if a beta distribution is assumed for θ , it can be proved (see Appendix A, or [Johnson and Kotz \(1977\)](#) for a wider discussion) that the marginal distribution of a given set of outcomes is approximately the same as the one obtained according to the Polya scheme described above.

The notion of frequency was briefly raised because it is commonly found among many scientists, and it is useful to clarify the way in which it is related, but *not equivalent*, to probability. A forensic scientist in fact may often encounter occasions where probability, which refers to personal beliefs, and chance, which refers to the probability of success in a Bernoulli series, are numerically the same ([Lindley, 1991](#)). The evaluation of scientific evidence in Court is the expression of one's personal degree of conviction: this often requires a combination of data on the occurrence of target features, summarised for example in terms of relative frequencies, together with a personal knowledge of circumstances for a particular case. Related to this is a shortcut widely observed among forensic scientists that equates beliefs with relative frequencies (see [Lindley \(2002\)](#) for discussion on this observation in other contexts). This gives rise to confusion between *frequency*, a term which refers to data, and *probability*, which refers to personal beliefs. Clearly, any probability judgement referring to a particular case, even when thought of in a frequency format, has a component based on personal knowledge. If probabilities are given a subjective interpretation, the relative frequencies (i.e., the population proportions) are known, and each individual in the population has the same probability of possessing a given characteristic of interest, that is observations are realizations of independent and identically distributed Bernoulli trials given probability of success θ , then relative frequencies determine the individual probabilities. For a recent critical discussion of the notion of frequency in relation to the rarity of particular genetic traits, see also [Brenner \(2014\)](#).

With these considerations in mind, the next Section will address an issue that arises in discussions about the assignment of the likelihood ratio in forensic science applications, in particular the idea held in informal discussions by some quarters in the field that the likelihood ratio could be 'estimated'.

The question is whether the Bayesian learning about the population parameter of interest does entitle one or not to claim a Bayesian inference about the likelihood ratio.

4. ESTIMATION VERSUS ASSIGNMENT

As for many scientists in various other disciplines, parameter estimation is also of interest to forensic scientists. It is the topic that deals with the question of how one can learn from experience to draw inferences about the value of a population parameter. In essence, this notion refers to the process of using data to derive the value for a population parameter, such as a mean, a variance or a proportion. A main aspect of emphasis here is that the procedure refers to a real underlying quantity, such as a weight, a length, or a population proportion. A population proportion, for example, actually exists, but it may be unknown to scientists, hence they seek to estimate it. On a conceptual account, the very existence of an actual value of the parameter of interest represents a necessary requirement for the use of the terms ‘estimate’ and ‘estimation’. The importance of this understanding will be further clarified in Section 5.

Return to the typical case illustrated in the Section 2, where a stain was recovered at the crime scene and the laboratory reports a match between the genetic profiles of recovered and control material, denoted as x_r and x_c , respectively. The evidence E is x_r , the profile obtained from the stain at the crime scene and whose source is unknown. The control material is the profile x_c obtained from the suspect and whose source is known. The likelihood ratio reduces to 1 divided by a conditional probability: for simplicity, consider it as the probability γ of observing an individual with the observed genetic profile in the relevant population. In some situations the population might be quite narrowly defined, in other cases less so, but in nearly every case information about all the members of this population will not be available. In practice, this quantity is not available, which is why it is widely considered (‘estimated’) in terms of the allelic relative frequency f found in an available database of individuals. These individuals are able to be assigned to two particular mutually exclusive and exhaustive categories, those who possess a given (‘target’) DNA sequence, and those who possess another DNA sequence, different from the target one. This value f is strictly data-based and has no elements of uncertainty. The observed relative frequency f must not be confused with the limiting frequency about which there is uncertainty and about which we express beliefs through the introduction of a beta distribution, say $\theta \sim Be(\alpha, \beta)$ where $Be(\alpha, \beta)$ denotes the beta distribution with probability density function

$$f(\theta | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}.$$

Let $\mathbf{x} = (x_1, \dots, x_n)$ denote the data available at the expert’s disposal, where x_i denotes an observation made on the i -th individual, an individual drawn randomly from the population of interest and indicates whether the target sequence of DNA is observed or not and n is the number of people in the sample. The source of the target sequence is assumed not to be in the sample. The variable x_i is a binary variable with $x_i = 1$ if the DNA of individual i matches the DNA of the target and $= 0$ otherwise. Thus \mathbf{x} is a sequence of trials each of which has a binary outcome. Whenever no subpopulation structure is considered, the observed outcomes, $x_i, i = 1, \dots, n$, can be treated as realisations of independent and identically distributed Bernoulli trials given the probability of success θ . Let $s = \sum_{i=1}^n x_i$ represent the observed number of individuals having the target DNA sequence. The expert’s uncertainty about the population parameter θ can then

be described by a beta posterior distribution $Be(\alpha + s + 1; \beta + n - s)$, where the parameters (α, β) of the prior distribution are updated according to well-established results. The prior parameter α is increased by the number of profiles $(s + 1)$ that match the recovered profile, including that of the suspect (control) profile. The prior parameter β is increased by the number of profiles $(n - s)$ that do not match the recovered profile. In the case described so far, the updated conditional probability for the denominator is given by $Pr(E | H_d, x_c, \mathbf{x})$

$$\begin{aligned} Pr(E | H_d, x_c, \mathbf{x}) &= \int_{\theta} Pr(E | H_d, x_c, \mathbf{x}, \theta) Pr(\theta | H_d, x_c, \mathbf{x}) d\theta \\ &= \int_{\theta} \theta Be(\alpha + s + 1, \beta + n - s) d\theta \\ &= \frac{\alpha + s + 1}{\alpha + \beta + n + 1} \end{aligned}$$

where $Pr(E | H_d, x_c, \mathbf{x}, \theta) = Pr(x_r | H_d, x_c, \mathbf{x}, \theta) = \theta$. Note that $Pr(E | H_d, x_c, \mathbf{x})$ is given by the posterior mean of a beta-distributed random variable θ . By rearranging terms in the previous equation one may observe that

$$Pr(E | H_d, x_c, \mathbf{x}) = \left(\frac{\alpha}{\alpha + \beta} \times \frac{\alpha + \beta}{\alpha + \beta + n + 1} \right) + \left(\frac{s + 1}{n} \times \frac{n}{\alpha + \beta + n + 1} \right),$$

that is the posterior mean can be expressed as a linear combination between the prior mean $\frac{\alpha}{\alpha + \beta}$ and the observed relative frequency $\frac{s+1}{n}$, with weights such that the larger the available data, the smaller influence will be ruled out by the prior distribution.

It is emphasised that the association of a probability distribution with parameter θ is to describe the *uncertainty* about the true unknown value of θ . The association is not the representation of its *variability*: θ is a fixed quantity, it is a number, it may be difficult to assess but it is a number and therefore does not have variability. Expression of $Pr(E | H_d, x_c, \mathbf{x})$ as the posterior mean of a beta distribution reflects posterior uncertainty about θ . This formulation of $Pr(E | H_d, x_c, \mathbf{x})$ is the origin of what is occasionally and loosely termed as '*uncertainty about the likelihood ratio*'.

Some more terminology is required to emphasise the difference between variability and uncertainty. A numerical value determined for a LR from given training data, control and recovered data will be known as an *assignment*⁶. The value calculated for the LR in such a situation is an objective assignment. Personal subjectivity only enters the enumeration through the choice of the prior parameters *e.g.* α and β . Two people with the same subjective choices of prior parameters, control and recovered data, training data and models will obtain the same value of the LR. The true value of the LR is unknown. The value I assign to the LR is my best assignment given the data, values for prior parameters and background information, I . There is uncertainty about my assignment because the true value is unknown. Uncertainty about a true unknown value of LR must not be confused with the notion of *variability* of the likelihood ratio. The difference is illustrated in Section 5.

⁶ This terminology is introduced to distinguish the procedure from that of estimation, with its statistical implication of variability and interval estimation.

The questions left open at the end of Section 3 can now be answered: does Bayesian learning about θ entitle one to claim a Bayesian learning process about the likelihood ratio? Is it possible to *estimate* a likelihood ratio? If we can agree that, following the above discussion, the scientist's probability provided for the denominator of the likelihood ratio depends on the posterior uncertainty about parameter θ , and this parameter refers directly to a real state of nature (*i.e.*, a population proportion), one can consider the choice of a parameter value (*i.e.*, the posterior mean) as a necessary preliminary decision for an assignment of a likelihood ratio. This reflects the idea that the problem of likelihood ratio assignment is broken down into local problems which can be analysed and discussed in a defensible way. However, there is no meaningful state of nature equivalent for the likelihood ratio in its entirety, as it is given by a ratio of two conditional probabilities. These probabilities may be hard to specify, but they are given by single values, and so is the likelihood ratio.

Another aspect related to parameter estimation is the fact that, generally, 'point estimation' is taken as some process of arriving at an estimate without regard to its precision. Interval estimation, where the precision of the estimate is to some extent taken into account, deals with this given aspect. The idea is to build a so-called credible interval, that specifies a range of values in which the parameter of interest, say θ , is located with a given probability. This allows one to specify that θ has, for example, a $100(1 - p)\%$ probability to be in a range defined by a given pair of a lower and an upper bound. It is tempting to consider the same arguments in the scenario under discussion: since there is uncertainty about the proportion of the population that has the target DNA sequence, and this uncertainty may be transposed to the likelihood ratio, with the ensuing suggestion to build some sort of interval for this latter. Curran (2005), for example, considered simulation techniques to build the sampling distribution of the likelihood ratio from which sample quantiles were extracted to act as endpoints of such an interval. However, even if it were sensible to determine an interval, it is not clear how such an interval should act as a multiplication factor for the prior odds in favour of a prosecution proposition, held by a recipient of expert information. Some authors advocate the presentation of a lower bound in order to avoid confusion that may arise from presenting a range of values to the Court. This represents an arbitrary choice that may be difficult to justify. With the current scenario, a reported lower bound would equate the selection of a value for θ characterised by a small probability density, with no reasonable justification, apart from grounds of simplicity.

Clearly, a forensic scientist may deal with different contexts, and the likelihood ratio may involve probability densities, as in (2.2). In some cases data present regular characteristics that can reasonably be described using standard parametric models. For example, a natural choice for data that are approximately unimodal, symmetric and bell-shaped is the Normal distribution. In this case, the task of the forensic scientist reduces to the estimation of the distribution's parameters under propositions H_p and H_d . However, forensic scientists often deal with situations where data do not have such regular characteristics that make it suitable to use a standard distribution. In these cases, a probability density function can be estimated from available data, for example with the use of kernel density estimation. Kernel density estimation is not a novel approach in forensic science, and several

applications can be found (see, *e.g.* Aitken and Taroni (2004)). In such a context, it can be said that the forensic scientist will use the background information to estimate a probability density function for each of the competing propositions. These estimates, given the available control and recovered measurements (x, y) , will enable computation of the value of the numerator and the denominator in (2.2) and hence the assignment of a value to the likelihood ratio. The likelihood ratio for the evaluation of a particular piece of evidence in a particular case is thus represented by a number.

5. UNCERTAINTY AND VARIABILITY

There is uncertainty about the value of the likelihood ratio, as explained in Section 4. However, it is inappropriate to represent this uncertainty as an interval estimate of the true value of the LR.

There is, however, a possible use of the term ‘distribution’, or *likelihood ratio distribution*, but its interpretation differs fundamentally from what has been discussed so far in that it is used in pre-assessment. Suppose that given two propositions of interest under which the findings should be evaluated, the scientist is able to assign a value to a likelihood ratio. Imagine, for sake of illustration, that a value of 10 is assigned. The scientist may thus report that this result slightly supports the prosecution’s case, by a factor of 10. Such a result may give rise to questions of the following kind: ‘How much credit should we owe to this result?’, ‘To what extent can we rely upon this result?’.

To approach this topic, scientists may consider related questions, that are more readily tractable, in particular: ‘How often may a forensic scientist obtain such a likelihood ratio for observations and findings that actually come from a population characterizing the prosecution’s case, H_p (*e.g.*, when items from the same source are analysed and compared)?’, ‘How often may a forensic scientist obtain such a likelihood ratio for observations and findings in a situation as specified by the alternative proposition assumed by the defence case, H_d (*e.g.*, when the compared materials come from different sources)?’. Note that the answers to these questions are not of particular relevance for the evaluation of evidence in a particular case. An answer to the latter question in a particular case is presented as a *p*-value in Dørrum *et al.* (2014) but such a use is subject to the same criticisms as have been made for the use of interval estimation.

To answer such questions, the scientist needs information about the likelihood ratio distribution, independent of the observations made in a given case. Such information is obtainable before findings are made. Several pairs of observations for each setting of interest are generated, and for every pair a likelihood ratio is determined. If this procedure is performed a certain number of times, that is for many pairs of observations under each setting of interest the forensic scientist obtains a sampling distribution of the likelihood ratio for each setting. In many situations, such data can be compiled through practical experiments because target materials and substances are easily available. As an example, consider the practical case mentioned in Section 2 where glass fragments are recovered at the crime scene, and continuous measurements on the recovered and control materials allow one to assign a given value to the likelihood ratio: how may the importance of this assignment be assessed? A likelihood ratio distribution may help to quantify how often a likelihood ratio taking values in a given range (containing the

one characterising the case at hand) can be obtained when the analysed findings come from the same source. Starting from a database containing measurements of glass fragments originating from several windows, pairs of fragments coming from the same and or from different windows are sampled and the likelihood ratio for each pair is computed. These results provide a distribution among several allocations of likelihood ratio results (Aitken and Lucy, 2004). Another possibility is to generate data for each of the target settings by simulations. Such an approach may be acceptable, for instance, in applications where the distribution of source features is well known and can be used to generate particular values by simulation. Consider for example the case of kinship determination.

Given appropriate population data, one can create databases of pairs of individuals as suggested by Triggs and Buckleton (2002). For example, one can generate a database of a large number of full siblings on the one hand, and a database of a large number of unrelated individuals on the other hand. For each pair of individuals coming from each database separately, a likelihood ratio is determined. In this way, two distributions of likelihood ratio values are obtained, one for the populations of pairs of siblings and one for the population of unrelated pairs. Such distributional information may be valuable in consideration of questions during pre-assessment, such as: *Is it possible to obtain a value supporting the hypotheses of interest in this scenario?*

The data thus generated are helpful for the assessment of what may be called the reliability or robustness of an analytical methodology (say, a 16 DNA loci kit) with an associated likelihood ratio. The scientist may focus, for example, on how many times a particular likelihood ratio – obtained for a particular setting (*e.g.*, same or different source) – points in the wrong direction (*i.e.*, supports the first proposition instead of the second, or vice-versa). This informs about the potential of misleading evidence. For the sake of illustration, consider the case of kinship determination. Simulation studies show that likelihood ratio distributions tend to overlap, and that there are values of the likelihood ratio that appear to be reasonable under both propositions of interest (Taroni *et al.*, 2005). Such simulation studies thus allow scientists to give a general assessment of false negatives and positives for a given setting, and assess the discriminative capacity of a particular examination and evaluation procedure with respect to particular propositions of interest. Note that several statistics coping with misleading evidence and performance measures for likelihood ratios are available in literature. Corradi and Ricciardi (2013), for example, propose a probability distribution for the likelihood ratio conditional on competing hypotheses in a kinship identification scenario. However, these pre-assessment investigations should not distract one's attention from the fact that, for a given case assessment, a comparison is evaluated in terms of a single value of likelihood ratio or Bayes factor.

6. DISCUSSION AND CONCLUSION

Throughout the previous sections, the main argument was directed to point out the rationale and properties underlying some key terms that are regularly invoked in the context of probabilistic evaluative procedures in forensic science. In combination, these aspects point out that – in the main – it is both misleading and non-sensical to talk in terms of ‘estimation’ with respect to probabilities, Bayes factors and likelihood ratios. A main motivation for presenting this dis-

cussion is that scientific communications and forensic commentators regularly invoke positions that are in conflict with this understanding. Examples include statements such as '*there is uncertainty about the probability to be used*', or '*there is uncertainty about the value the likelihood ratio*', which appears to prompt some scientists to elaborate ways to construct intervals or distributions over probabilities and likelihood ratios.

Such attempts must be considered spurious in at least two respects. First, probability is a measure for uncertainty about the truth of a proposition, or about the actual value an unknown quantity may assume (say, a population proportion). By specifying probabilities/densities for various possible values of the target quantity one obtains what is called a probability distribution. One such example is the beta probability density that was introduced in Section 4 with reference to a particular hypothetical forensic scenario of interest. However, this does not allow to conclude that a given probability, say the probability $Pr(E | H_d)$ to find a person in the general population that possesses the target DNA sequence , should itself have a distribution. That would mean to confuse the probability with the unknown proportion of individuals possessing a given characteristic, and place a probability on a probability, which would result in an infinite regression, and would not solve the problem (de Finetti, 1976). Hence the title 'The dismissal of the illusion of uncertainty' for this paper.

Second, one must not confuse the difficulty of the *problem of measurement*, with the measuring device (probability) itself. This is not to ignore that the specification, in the first place, of a probability for a possible value of an unknown quantity (or, more generally, an event of interest) may be a difficult task. However, nothing will be gained if a particular expression for uncertainty, in terms of a probability, is itself obscured or blurred by an additional level of uncertainty. This same argument applies to likelihood ratios as a measure of probative value. These distinctions are subtle, but fundamental. In everyday language such distinctions are often coarse and these aspects are ignored. Scientific accounts cannot credibly do so.

This is also not to deny that, on a purely technical account, one can introduce additional mathematical concepts for the evaluation of evidence. However, such an introduction does not provide release from the obligation to provide a conceptually sound, contextually consistent and operationally meaningful rationale for such concepts. In practical matters of forensic science and law, it is important for there to be agreement on the understanding of the first principles. Technical matters should only be introduced once the objective is clearly defined, otherwise they will merely serve to obscure any impasse there may be on foundational matters.

The provision of probabilities and likelihood ratios can be a challenging issue, and the requirement to assign a single value can be perceived as difficult, or even intimidating. However, with these provisions and requirements, the difficulty is not one of uncertainty, because one is uncertain only about the real event, proposition or parameter in the first place, *nor* about how one is thinking about their truth or otherwise. The issue, instead, is more properly termed one of 'precision'. Thus, one way to approach the difficulty of measurement and the communication of uncertainty, and hence the probative value, could be to avoid the use of numbers, as their use may suggest an unrealistically high precision for the value of a

probability or likelihood ratio to several significant figures, for example. Instead, scientists should devote their efforts to the elicitation and communication of results that they can present – within a personal and decision-theoretic perspective – as their best representation of the value of the issue under consideration; *i.e.*, as a probability or a likelihood ratio. Associated with these efforts in individual cases, there is a general need for the instruction of recipients of expert information (*e.g.*, jurists) in the nature of probability, the importance of an understanding of it and its proper use in dealing with uncertainty.

ACKNOWLEDGEMENTS

This research was supported by the Swiss National Science Foundation, grant numbers 100012-144227 and IZK0Z1-155210.

REFERENCES

Aitken, C. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, **53**, 109–122.

Aitken, C. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley and Sons, Chichester, second edition.

Alberink, I. and Bolck, A. (2008). Obtaining confidence intervals and Likelihood Ratios for body height estimations in images. *Forensic Science International*, **177**, 228–237.

Aven, T. and Reniers, G. (2013). How to define and interpret a probability in a risk and safety setting. *Safety Science*, **51**, 223–231.

Brenner, C. H. (2014). Understanding Y haplotype matching probability. *Forensic Science International: Genetics*, **8**, 233–243.

Corradi, F. and Ricciardi, F. (2013). Evaluation of kinship identification systems based on short tandem repeat dna profiles. *Applied Statistics*, **62**, 649–668.

Curran, J. (2005). An introduction to Bayesian credible intervals for sampling error in dna profiles. *Law, Probability & Risk*, **4**, 115–126.

de Finetti, B. (1974). *Theory of Probability, A critical introductory treatment, Volume 1*. John Wiley & Sons, London.

de Finetti, B. (1976). *Probability, Induction and Statistics*. John Wiley & Sons, London.

Dørum, G., Bleka, Ø., Gill, P., Haned, H., Snipen, L., Sæbø, S., and Egeland, T. (2014). Exact computation of the distribution of likelihood ratios with forensic applications. *Forensic Science International: Genetics*, **9**, 93–101.

Hacking, I. (1975). *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge University Press, Cambridge.

Hays, W. L. and Winkler, R. L. (1970). *Statistics: Probability, Inference, and Decision, Volume 1*. International Series in Decision Processes. Holt, Rinehart and Winston, Inc., New York.

Jeffreys, H. (1931). *Scientific inference*. Cambridge University Press, Cambridge.

Johnson, N. and Kotz, S. (1977). *Urn models and their application*. John Wiley & Sons, New York.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Laplace, P.S. (1814). *A Philosophical Essay on Probabilities (English edition 1951)*. Dover Publications, Inc., New York.

Lindley, D. (2006). *Understanding Uncertainty*. John Wiley & Sons, Hoboken.

Lindley, D. V. (1991). Probability. In C. G. G. Aitken and D. A. Stoney, editors, *The use of statistics in forensic science*, pages 27–50. Ellis Horwood, New York.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, **49**, 293–337.

Lindley, D. V. (2002). Seeing and doing: the concept of causation. *International Statistical Review*, **70**, 191–214.

Simmross, U. (2014). Appraisal of scientific evidence in criminal justice systems: on winds of change and coexisting formats. *Law, Probability and Risk*, **13**, 105–115.

Stoel, R.D. and Sjerps, M. (2012). Interpretation of forensic evidence. In S. Roeser, R. Hillerbrand, P. Saad and M. Peterson, editors, *Handbook of risk theory. Epistemology, decision theory, ethics and social implications of risk*, pages 135–158. Springer, The Netherlands.

Taroni, F., Bozza, S., and Aitken, C. (2005). Decision analysis in forensic science. *Journal of forensic sciences*, **50**, 894–905.

Triggs, C. and Buckleton, J. (2002). Logical implications of applying the principles of population genetics to the interpretation of DNA profiling evidence. *Forensic Science International*, **128**, 108–114.

APPENDIX A: APPENDIX

Consider an urn composed by w white balls and b black balls, and the Polya learning scheme illustrated in Section 3, according to which after each drawing the chosen ball is returned together with a ball of the same colour. Define a variable X_i that takes value $x_i = 1$ if a white ball is observed at the i -th extraction, and value $x_i = 0$ viceversa. Then, the probability to observe a white ball at the second extraction can be revised as:

$$Pr(X_2 = x_2 | X_1 = x_1) = \frac{w + x_1}{w + b + 1}.$$

Similarly, at the i -th iteration the probability to extract a white ball is easily computed as

$$(A.1) \quad Pr(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) = \frac{w + \sum_{j=1}^{i-1} x_j}{w + b + i - 1}.$$

Consider a sequence of n draws: the joint probability $Pr(X_1 = x_1, \dots, X_n = x_n)$ can be simply obtained by multiplying conditional probabilities in (A.1) for $i = 1, \dots, n$. For example given $n = 5$ draws, the probability of the generic set of outcomes $WBBWW$ can be computed as

$$Pr(X_1 = 1) \times Pr(X_2 = 0 | X_1 = 1) \times \dots \times Pr(X_5 = 1 | X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1) = \\ \frac{w}{w+b} \times \frac{b}{w+b+1} \times \dots \times \frac{w+2}{w+b+4} = \frac{w(w+1)(w+2)b(b+1)}{(w+b) \dots (w+b+4)} = \frac{w^{[3]}b^{[5-3]}}{(w+b)^{[5]}},$$

where $w^{[k]} = w(w+1) \dots (w+k-1)$. In the same way, given n draws there one can compute the probability to observe a total number s of white balls and a total number $(n-s)$ of black balls. Consider a variable Y that counts the number of white balls that have been observed after n draws. Then one has

$$(A.2) \quad Pr(Y = s) = \binom{n}{s} \frac{w^{[s]}b^{[n-s]}}{(w+b)^{[n]}},$$

where $\binom{n}{s}$ represents the number of sequences that may be given by s white balls and $(n-s)$ black balls.

Suppose now variables X_i are independent, that is suppose they can be treated as independent Bernoulli trials with a probability to observe a success (i.e., to draw a white ball), given by θ , and a probability to observe a failure (i.e., to draw a black ball) given by $(1-\theta)$. Then, given θ , the probability of any particular sequence containing s white balls and $(n-s)$ black balls is

$$Pr(Y = s | \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}.$$

If a Beta distribution is introduced on θ , say $\theta \sim Be(\alpha, \beta)$, the marginal probability to observe any sequence of outcomes with s successes and $n - s$ failures can be easily verified to be a Beta-Binomial distribution with parameters n, α, β , that is

$$(A.3) \quad Pr(Y = s) = \binom{n}{s} \frac{\Gamma(\alpha + s)\Gamma(\beta + n - s)}{\Gamma(\alpha + \beta + n)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

Rearranging terms in (A.3), one can observe that the marginal distribution is the same as that obtained in (A.2) with the Polya sequence. In fact, recalling that for any integer k the Gamma function satisfies the relationship $\Gamma(k) = (k - 1)!$, then

$$\begin{aligned} Pr(Y = s) &= \binom{n}{s} \frac{(\alpha + s - 1)!(\beta + n - s - 1)!}{(\alpha + \beta + n - 1)!} \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} \\ &= \binom{n}{s} \frac{\alpha(\alpha + 1) \cdots (\alpha + s - 1)\beta(\beta + 1) \cdots (\beta + n - s - 1)}{(\alpha + \beta)(\alpha + \beta + 1) \cdots (\alpha + \beta + n - 1)} \\ &= \binom{n}{s} \frac{\alpha^{[s]}\beta^{[n-s]}}{(\alpha + \beta)^{[n]}}. \end{aligned}$$

The described learning scheme can be regarded as being generated by a sequence of independent and identically distributed Bernoulli trials, in which the parameter θ has a beta distribution. Thus, if the Polya learning scheme is considered to revise prior beliefs about the composition of the urn, it is the uncertainty about the urn composition θ that is characterised by a beta distribution, while subsequent draws are considered as independent and identically distributed Bernoulli trials. See [Johnson and Kotz \(1977\)](#).