Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach)

Giulia Cereda,[1*,2]

[1]University of Lausanne, Faculty of Law, Criminal Justice and Public Administration
[2]Leiden University, Mathematical Institute

*To whom correspondence should be addressed; E-mail: giulia.cereda7@gmail.com.

The likelihood ratio (LR) measures the relative weight of forensic data regarding two hypotheses. Several levels of uncertainty arise if frequentist methods are chosen for its assessment: the assumed population model only approximates the true one and its parameters are estimated through a limited database. Moreover, it may be wise to discard part of data, especially that only indirectly related to the hypotheses. Different reductions define different LRs. Therefore, it is more sensible to talk about "a" LR instead of "the" LR, and the error involved in the estimation should be quantified. Two frequentist methods are proposed in the light of these points for the 'rare type match problem', that is when a match between the perpetrator's and the suspect's DNA profile, never observed before in the database of reference, is to be evaluated.

Key words: Evidence evaluation, frequentist approach, likelihood ratio, rare type match, uncertainty, Y chromosome STR.

1 Introduction

One of the main challenges of forensic science is to evaluate how much some evidence can be helpful to discriminate between hypotheses of interest. For instance, a typical piece of evidence may be a DNA trace which is found at the crime scene and whose profile matches a known suspect's DNA profile. A couple of mutually exclusive hypotheses is typically defined, of the kind of 'the crime stain came from the suspect' (h_p) and 'the crime stain came from an unknown donor' (h_d). The largely accepted method to perform this evaluation is the calculation of the *likelihood ratio*, a statistic that expresses the relative

plausibility of the observations under the two hypotheses (Robertson and Vignaux, 1995; Evett and Weir, 1998; Aitken and Taroni, 2004; Balding, 2005; Steele and Balding, 2014).

The definition of the likelihood ratio depends on whether a Bayesian or a frequentist approach is chosen. In the Bayesian context, after a couple of hypotheses is given, the likelihood ratio is defined as

$$\mathrm{LR} = \frac{\Pr(D = d \mid H = h_p)}{\Pr(D = d \mid H = h_d)}, \tag{1}$$

where Pr is the Bayesian probability, reflecting the expert's belief on the joint distribution of the random variables of the model, namely D (representing the data), H (representing the hypotheses), and Θ (the nuisance parameter(s)).

On the other hand, in a frequentist context, the nuisance parameter θ and the hypotheses h are considered to be fixed (unknown) quantities. The frequentist probability (here denoted as $\mathcal{P}\nabla$) can be expressed in terms of the Bayesian Pr, in the following way: $\mathcal{P}\nabla_\theta(\cdot \mid h) := \Pr(\cdot \mid \Theta = \theta, H = h),\ \forall h$. The frequentist likelihood ratio can be thus expressed as

$$\mathit{Lr}_\theta = \frac{\mathcal{P}\nabla_\theta(D = d \mid h_p)}{\mathcal{P}\nabla_\theta(D = d \mid h_d)}. \tag{2}$$

It is important to consider that different reductions of the data D can be carried out, each corresponding to a different frequentist likelihood ratio. Moreover, unless we choose nonparametric solutions, a model choice is also performed, and there are often parameters to be estimated. Hence, two further levels of uncertainty have to be added to the initial uncertainty regarding which hypothesis is the true one.

The main aim of this paper is to provide the message that, if a frequentist approach is chosen and an estimation is needed, (i) it is more sensible to talk about "a" likelihood ratio instead of "the" likelihood ratio, and (ii) a quantification of the error involved in the estimation of the likelihood ratio is to be provided along with the estimated value.

It is believed in the forensic field that the use of frequentist methods to assess the likelihood ratio is not coherent, since the likelihood ratio has to be used within the Bayes' theorem context, as the way to update prior odds to posterior odds. However, frequentists may be interested as well in the likelihood ratio, seen as a tool to measure the evidential value of data, independently of the Bayes' theorem. Moreover, literature presents many approaches to calculate the likelihood ratio, wrongly defined as Bayesian, which in fact plug in Bayes estimates into a likelihood ratio defined in a frequentist way (for a discussion, see Cereda, 2015). We thus believed that it is important to study and discuss the two approaches (the Bayesian and the frequentist) separately, in order to define coherent methodologies and avoid unnecessary hybrid methods. This is done in Section 2.

In forensic science, a very challenging problem is the so-called *rare type match*, the situation in which there is a match between the characteristics of some recovered material

and the corresponding characteristics of the control material, but these characteristics have not been observed yet in previously collected samples (i.e., they do not occur in any existing database of interest for the case). This constitutes a problem because of the presence of a nuisance parameter that is (related to) the proportion of individuals (or items) in possess of the matching characteristic in a reference population: this proportion is, in standard frequentist practice, estimated using the relative frequency of the characteristic in a previously collected database. Thus, in case of rare type match, there's the need for different solutions.

This paper discusses two frequentist methods to provide a likelihood ratio in the rare type match case, based respectively on the parametric discrete Laplace method (Andersen et al., 2013b), and on a generalization of the nonparametric Good-Turing estimator (Good, 1953). The latter looks similar to Brenner's 'κ-method' (Brenner, 2010), but is different inasmuch it does not need any assumption and provides two different frequencies, one for the prosecution's and one for the defense's point of view. We plan to compare the two methods in a future paper.

More specifically, these two methods are here proposed as an answer to the problem of the rare Y-STR haplotype match: the situation in which the matching (and previously unseen) characteristic is a Y-STR profile. Each of the two methods is analyzed in the light of points (i) and (ii) discussed above, by carefully specifying the data reduction, the chosen probability model, and with a discussion on the different levels of error involved in the estimations.

Sections 3 and 4 draw out in depth the rationale behind points (i) and (ii) above, Section 5 describes the paradigmatic example of the rare Y-STR haplotype match problem, to which we will apply the discrete Laplace method (Section 6), and the Generalized-Good method (Section 7) according to the guidelines exposed in the opening sections.

2 Bayesian versus frequentist approach to likelihood ratio assessment

The task of a forensic statistician is to measure the extent to which some given data favors one hypothesis instead of the other. For instance, the data at disposal may consist of a DNA trace found at the crime scene which matches a suspect's DNA profile, and of a database of collected DNA profiles from a reference population or past cases. This is a paradigmatic example to which, from now on, we will refer generically as "the DNA example". The prosecution and defense hypotheses are usually of the kind "the trace has been left by the suspect" (h_p) and "the trace has been left by an unknown person" (h_d). Denote with $h \in \{h_d, h_p\}$ the unknown true hypothesis, and with θ the nuisance parameter involved in the assessment of the likelihood ratio. In the DNA example, the vector made

of all the DNA frequencies can be thought of as the nuisance parameter θ. Notice that there is a difference between h and θ: one (h) is the parameter which we 'test' through the likelihood ratio, the other (θ) is a nuisance parameter involved in the likelihood ratio assessment. It is often possible to split the data D into E, evidence directly related to the crime, and B, additional information not related to the crime and only pertaining to the nuisance parameter θ. In the DNA example, we can take as E the couple of matching profiles (that of the trace and that of the suspect) and as B the database of reference. D, E, and B can be regarded as random variables, such that $D = (E, B)$.

Bayesian and frequentist methods differ in how they consider the parameters θ and h. In a Bayesian context they are modelled through random variables Θ and H, which are given prior distributions $p(\theta)$ and $p(h)$. Frequentists consider them as fixed (i.e., without distribution) unknown quantities. Regardless of the type of approach which is chosen, some model assumptions concerning E and B, θ and h can be made:

a. The distribution of B given h and θ, only depends on θ.

b. B is independent of E, given h and θ.

In the DNA example, condition **a** holds if for instance the database is collected before the crime, since the sampling mechanism to obtain the database of reference is independent of which hypothesis is correct. Condition **b** holds if the suspect has been found on the ground of different evidence that has nothing to do with DNA.

2.1 The Bayesian approach

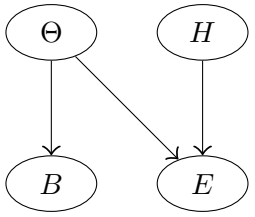


Figure 1: Bayesian network representing the dependency relations between E (evidence of the case), B (background data), Θ (nuisance parameter), and H (hypotheses of interest).

A full Bayesian model is defined by giving the prior joint probability distribution for all the random variables of the model (here E, B, H and Θ). It can be represented by the Bayesian network of Figure 1, which is in turn equivalent to the following Bayesian reformulation of conditions **a**, and **b**, with a third additional condition:

Bayesian a. B is conditionally independent of H given Θ.

Bayesian b. B is conditionally independent of E given Θ and H.

Bayesian c. Θ is unconditionally independent of H.

Condition **Bayesian c** is guaranteed for instance if prior beliefs on θ and on h are assessed by people with different responsibilities and tasks: a judge for h and a forensic DNA expert (or a statistician) for θ. The joint prior can be factorized as follows, by looking at the structure of the Bayesian network or, equivalently, using the three conditions above: $p(\theta, h, b, e) = p(\theta)p(h)p(b|\theta)p(e|\theta, h)$. By choosing a prior distribution for θ and h which reflects expert's beliefs, the Bayesian probability is an expression of the subjective credence of the experts. The distribution of all other variables given θ and h is defined by the structure of the model, and needs no subjective assessment.

The Bayesian likelihood ratio can be derived in the following way:

$$\begin{aligned}\mathrm{LR} =& \frac{\Pr(E = e, B = b \mid H = h_p)}{\Pr(E = e, B = e \mid H = h_d)} = \frac{\Pr(E = e \mid B = b, H = h_p)}{\Pr(E = e \mid B = b, H = h_d)} = \frac{\int p(e \mid b, h_p, \theta)\, p(\theta \mid b, h_p) \mathrm{d}\theta}{\int p(e \mid b, h_d, \theta) p(\theta \mid b, h_d) \mathrm{d}\theta} \\ =& \frac{\int \theta\, p(\theta \mid b)\, \mathrm{d}\theta}{\int \theta^2\, p(\theta \mid b)\, \mathrm{d}\theta} = \frac{\mathbb{E}(\Theta \mid B = b)}{\mathbb{E}(\Theta^2 \mid B = b)}.\end{aligned}$$

Some simplifications have been carried out because of conditions **a**, **b**, and **c**. Since it is possible to marginalize out over all values of Θ, using its distribution, there's no need to estimate the likelihood ratio, or to account for uncertainties, if a proper full Bayesian approach is chosen.

In the rest of the paper we only focus on frequentist methods to solve the rare Y-STR haplotype match problem, but a companion paper presents a similar study on Bayesian methods (Cereda, 2015).

2.2 The frequentist perspective

The difference between frequentist and Bayesian methods regards parameters h and θ: for a frequentist they are fixed quantities, whose values correspond to, respectively, the unknown true value of θ and the correct hypothesis. One can see frequentist models as Bayesian models where the distributions chosen for Θ and H give probability one to values θ and h, respectively. Also, one can express the frequentist probability $\mathscr{Pr}$ in terms of the Bayesian probability Pr, in the following way: $\mathscr{Pr}(\cdot \mid h) := \mathscr{Pr}_\theta(\cdot \mid h) = \Pr(\cdot | H = h, \Theta = \theta)$. For frequentist statisticians, there is a true, 'physical' probability which governs the situation at hand: according to the prosecution this true probability is $\mathscr{Pr}_\theta(\cdot \mid h_p)$, while according to the defense it is $\mathscr{Pr}_\theta(\cdot \mid h_d)$, with θ set to its true (unknown) value.

Conditions **a** and **b** can be rephrased, in a frequentist language as:

Frequentist a. $\mathscr{Pr}_\theta(B = b \mid h_p) = \mathscr{Pr}_\theta(B = b \mid h_d)$, for all θ and b.

Frequentist b. $\mathscr{Pr}_\theta(E = e \mid B = b, h) = \mathscr{Pr}_\theta(E = e \mid h)$, for all θ, h, e, and b.

It holds that:

$$\mathcal{L}r = \frac{\mathcal{P}r(D = d \mid h_p)}{\mathcal{P}r(D = d \mid h_d)} = \frac{\mathcal{P}r(E = e, B = b \mid h_p)}{\mathcal{P}r(E = e, B = b \mid h_d)} = \frac{\mathcal{P}r(E = e \mid B = b, h_p)}{\mathcal{P}r(E = e \mid B = b, h_d)} \frac{\mathcal{P}r(B = b \mid h_p)}{\mathcal{P}r(B = b \mid h_d)}.$$

The index θ has been omitted for ease of notation. Thanks to conditions **Frequentist a** and **b**, the likelihood ratio can be expressed as

$$\mathcal{L}r = \frac{\mathcal{P}r(E = e \mid h_p)}{\mathcal{P}r(E = e \mid h_d)}. \tag{3}$$

Even though the two alternative ways of writing the likelihood ratio expressed by equations (2) and (3) are theoretically different, and mean two different things, they have the same value. This implies that part of the information, namely B, is not useful to discriminate between the two hypotheses of interest. Stated otherwise, when knowing θ, B is irrelevant to determine the likelihood ratio, i.e. to decide about parameter h. However, it may play an important role in the estimation of parameter θ. For instance, getting back to the DNA example, the database (B) is often useful to estimate the frequencies of the different haplotypes.

Notice that, in order for (3) to hold, **b** can be modified to something less strong:

Frequentist b*. $\dfrac{\mathcal{P}r_\theta(E = e \mid B = b, h_p)}{\mathcal{P}r_\theta(E = e \mid B = b, h_d)} = \dfrac{\mathcal{P}r_\theta(E = e \mid h_p)}{\mathcal{P}r_\theta(E = e \mid h_d)}$ for all e, b, and θ.

which is equivalent to ask that updating the likelihood ratio for the observation of B to take into account the observation of E, does not change anything.

Furthermore, while conditions **a** and **b*** imply (3), the converse is not true. Formulation (3) is instead equivalent to a weaker condition, that is:

Frequentist c. $\mathcal{P}r_\theta(B = b \mid E = e, h_p) = \mathcal{P}r_\theta(B = b \mid E = e, h_d)$, for all θ.

This can be seen by the following alternative development of the likelihood ratio (θ omitted):

$$\mathcal{L}r = \frac{\mathcal{P}r(D = d \mid h_p)}{\mathcal{P}r(D = d \mid h_d)} = \frac{\mathcal{P}r(B = b \mid E = e, h_p)}{\mathcal{P}r(B = b \mid E = e, h_d)} \frac{\mathcal{P}r(E = e \mid h_p)}{\mathcal{P}r(E = e \mid h_d)} = \frac{\mathcal{P}r(E = e \mid h_p)}{\mathcal{P}r(E = e \mid h_d)}. \tag{4}$$

It follows that:

$$\mathbf{c} \Leftrightarrow \mathrm{LR} = \frac{\mathcal{P}r(E = e \mid h_p)}{\mathcal{P}r(E = e \mid h_d)}. \tag{5}$$

Notice that frequentists use a likelihood ratio $\mathcal{L}r_\theta$, which can be written in terms of the Bayesian LR as $\mathrm{LR}|\Theta = \theta$ (read "LR given θ"), and attempt to get close to θ by choosing some estimator $\widehat{\theta}$. This leads to the so-called *plug-in estimator* $\widehat{\mathcal{L}r_\theta} = \mathcal{L}r_{\widehat{\theta}} = \mathrm{LR}|(\Theta = \widehat{\theta})$. However, that's not the only option, as we will see for the method explained in Section 7.

It is important to notice that the frequentist approach may be represented by the same Bayesian network of Figure 1, where the states of nodes Θ and H are instantiated to particular values θ and h, respectively. This shows that actually the two approaches

don't disagree on the structure of the model regarding E and B. Only, Bayesians add ingredients to the model by allowing Θ and H to have a distribution. Stated otherwise, the Bayesian approach is defined by the very same frequentist conditions **a** and **b**, with the addition of condition **c** about the independence of Θ and H.

3 Data reduction

Let us denote with $\mathcal{D}$ all the data given to the expert in the form of a dossier, which he has to "translate" into a well-defined mathematical object. To evaluate the entirety of the data at the expert's disposal is often a delusion, from which the need for a reduction of $\mathcal{D}$ to something less informative, but of more feasible evaluation, which we denote as D. Often the database contains only information about a limited number of loci, and this implies that information about other loci of the crime stain can't be used. This constitutes already a first reduction of the data. Other kinds of reductions are performed in order to gain in terms of precision of the estimates. Especially in a situation with many nuisance parameters, it can be wise to discard the part of data which primarily tells us about the nuisance parameters, and only indirectly about the ultimate question of interest (i.e., which hypothesis is more likely to be true). In fact, it could be very wise to reduce the data $\mathcal{D}$ to a much smaller amount of information, because the likelihood ratio based on the data reduction is much more precisely estimated than one based on all data. However, there's a limit to this: the reduction of $\mathcal{D}$ into D comes with a cost: the stronger the reduction, the less the corresponding likelihood ratio value is discriminating of the two hypotheses, because less information is less powerful to that purpose. We have to make a compromise between a gain in terms of precision and a loss in strength of the evidence. This will be discussed more in detail in Section 8.

Once a particular reduction D has been defined, the frequentist likelihood ratio ($\mathcal{L}r$) can be defined as in (2). It is easy to understand that there isn't a unique way to reduce $\mathcal{D}$ and that each choice entails the definition of a different likelihood ratio. For instance, in the DNA example, one can consider a profile made of more or fewer loci. Another kind of reduction will be presented in Section 7. Different choices of $D \subsetneq \mathcal{D}$ lead to different likelihood ratios. Therefore, *it is better to refer to "a" likelihood ratio instead of to "the" likelihood ratio.* This was already stated in Dawid (2001), even though regarding hypotheses instead of data.

In the literature different choices of $D \subsetneq \mathcal{D}$ and 'Pr' are proposed, each corresponding to a different likelihood ratio to be estimated. These choices are often only implicit and one of the aims of this research is to make explicit the reduction which corresponds to two selected methods, by looking for the corresponding E and B.

4 Different levels of uncertainty

The likelihood ratio measures the relative strength of support given by the data to a hypothesis over an alternative. Clearly, it is useful when there is uncertainty about which of the two hypotheses is true (to be more precise, it may also be the case that none of the alternatives is correct, and the likelihood ratio continues to be meaningful). Along with this first basic initial uncertainty about the state of the affairs, two more levels of uncertainty arise in the attempt of calculating the likelihood ratio.

For a frequentist statistician, the likelihood ratio is a ratio of probabilities based usually on a model $\mathcal{M}$ which is at best only a good approximation to the truth. Moreover, they have to estimate parameters of that model by fitting it to the data in some database. Stated otherwise, after a particular choice of what is the data D to be considered, a population model is to be chosen and its parameters estimated using a limited sample. Some forensic literature (Morrison, 2010; Stoel and Sjerps, 2012; Curran et al., 2002; Curran, 2005) already pointed out the necessity for uncertainty assessment in the likelihood ratio estimation, even though they don't differentiate among levels. On the other hand, for a true Bayesian statistician there's no need for estimation, and no additional levels of uncertainty to be added, since the definition of the Bayesian Pr already includes not only beliefs about chances when picking people from that population, but also beliefs about parameters of the models, and beliefs about models.

This discussion may hopefully put an end to the debate as to whether it makes sense to talk about 'estimation' and 'uncertainty assessment' for the likelihood ratio. Stoel and Sjerps (2012) believe that "there are strong arguments for the notion of a "true" but unknown value of the likelihood ratio, given the relevant hypotheses and background information, and that it is important to consider the uncertainty. Ignoring the uncertainty can be strongly misleading". This point of view is also shared in Sjerps et al. (2016). On the other hand, to talk about estimation of the likelihood ratio is defined as "internally inconsistent, and hence misconceived" by Taroni et al. (2015). Both the points of view are correct, if correctly put into context: if a frequentist approach is chosen it is sensible to talk about 'estimation' and to deal with uncertainty assessment. On the other hand, in a full Bayesian context, they are misplaced.

Notice that Bayesianism is theoretically a very powerful interpretation of probability, but when it comes to applying Bayesian theory for practical purposes, even the most fervent Bayesian has to strike a balance between what is feasible and what is theoretically right and coherent according to the Bayesian perspective. He typically chooses a particular model as the correct one (as frequentists do), and/or he has to put convenient (rather than realistic) prior distributions on the parameters. Hence, whether Bayesian or frequentist approaches are chosen, the attempt to produce the likelihood ratio leads to several levels

of uncertainty which should be accounted for.

We will now discuss the two additional levels of uncertainty mentioned before. The second level of uncertainty pertains to the choice of a particular population model, which is only an approximation of the truth. This level of uncertainty may be reduced using nonparametric methods, that are based on fewer assumptions.

Given a particular population model, the third level of uncertainty pertains to the fact that the population parameters are not known. This may involve estimation of parameters (such as in the discrete Laplace method of Section 6) or the direct estimation of the probabilities of interest (as in the Generalized-Good method described in Section 7) and the quality of the estimates severely depends on the size of the available databases. This level of uncertainty pertains both to parametric and nonparametric methods.

The evidential value reported depends on all the levels of uncertainty which afflict the estimation of the likelihood ratio. Thus, it is of the utmost importance to report the likelihood ratio value along with (1) an explicit definition of which data D we want to evaluate through that likelihood ratio, and (2) a discussion (and if possible a quantification) of the levels of uncertainty that afflict the reported value.

4.1 Estimating the weight of evidence

Instead of estimating the likelihood ratio, it is more sensible to directly estimate its logarithm, sometimes called *relevance ratio* or *weight of evidence* (Good, 1950; Aitken et al., 1998; Aitken and Taroni, 2004). This is because the interpretation of the likelihood ratio values goes through orders of magnitude 10, and when a value is reported, it is important to control the relative error, rather than the absolute error. In fact, the first is meaningful in itself while the second depends on the particular value of the likelihood ratio. For the very same reasons why the verbal equivalent scale (Aitken et al., 1998) is based on logarithm. Furthermore, both the odds form of Bayes' theorem and the formula to combine likelihood ratios from independent pieces of evidence involve a multiplicative relationship that becomes a handier additive relation if logarithm is taken (Schum, 1994). Moreover, the logarithm helps in presenting large numbers in a compact way, of more easy comprehension, and it is symmetric with respect to prosecution's and defense's hypothesis: this may be useful if one wants to invert the weight of evidence to consider the defense's proposition (Aitken and Taroni, 2004).

5 The rare Y-STR haplotype problem

Consider the situation in which a piece of evidence is recovered at the crime scene, and a suspect turns out to have the same analyzed characteristics (for instance the same DNA profile) as the crime scene evidence. The prosecution claims that the suspect left the

evidence, defense claims that someone else (with the same DNA profile) left it. The capability of the match to discriminate between the competing hypotheses is evaluated by comparing how probable it is under each of the hypotheses. This depends on the proportion of individuals in possess of the same profile in the population of possible perpetrators: the rarer the profile the more the suspect is in trouble. This proportion is usually unknown, the only available data being a sample of DNA profiles from the population, in the form of a reference database. The *naive estimator* uses the relative frequency of the profile in the database as an estimate for θ. Problems arise when this frequency is 0, the so-called "rare type match". This problem is so substantial that it has been defined "the fundamental problem of forensic mathematics" by Brenner (2010). As an alternative to the empirical frequency estimator, one can use the *add-constant* estimators, which adds a constant to the count of each type, included the unseen ones. The most well known is the *add-one* estimator, due to Laplace (1814), and the *add-half* estimator of Krichevsky and Trofimov (1981). However, to use these methods one needs to know the number of possible unseen types and there are problems if this number is large compared to the sample size (see Gale and Church (1994) for additional discussion). Another possibility is the 'rule of three', proposed by Louis (1981). It states that $3/n$ is a good approximation of the 95% upper bound for the frequency, if n is the size of the database.

Of interest for this paper is the nonparametric *Good-Turing estimator* of Good (1953), based on an intuition on A. M. Turing. It is an estimator for the total unobserved probability mass which is based on the proportion of singletons in the sample. For a comparison between *add one* and *Good-Turing* estimator, see Orlitsky et al. (2003).

The *naive estimator* and the *Good-Turing estimator* are in some sense complementary (Anevski et al., 2013): the first gives a good estimate for the observed types and the second for the probability mass of the unobserved ones. Lastly, the *high profile estimator*, introduced by Orlitsky et al. (2004), extends the tail of the *naive estimator* to the region of unobserved types. This estimator has been improved by Anevski et al. (2013) that also provides the consistency proof.

The rare type match problem is common, for instance, in case a new kind of forensic evidence is involved, and for which the available database size is still limited. One example is the case of DIP-STR markers (e.g. Cereda et al., 2014). The same happens when Y-chromosome (or mitochondrial) DNA profiles are used: because of the lack of recombination involved when offspring DNA is generated from the DNA of the parents, each haplotype must be treated as a unit (the match probability can't be obtained by multiplication across loci) and the set of possible haplotypes is extremely large. As a consequence, most of the Y-STR haplotypes are not represented in the database.

In the rest of the paper, Y-STR data will be retained as an extreme but common and

important way in which the problem of assessing the evidential value of rare type match can arise. Literature provides some examples of approaches to evaluate it for the rare Y-STR haplotypes match: Egeland and Salas (2008), the κ method Brenner (2010, 2014), the coalescent theory method (Andersen et al., 2013a), the haplotype surveying method (Roewer et al., 2000; Krawczak, 2001; Willuweit et al., 2011), and the discrete Laplace method (Andersen et al., 2013b) (not directly proposed for the rare haplotype case but usable for that purpose). As already mentioned, Cereda (2015) discusses the full Bayesian approach to this problem.

Bayesian nonparametric estimators for the probability of observing a new type have been proposed by Tiwari and Tripathi (e.g. 1989); Lijoi et al. (e.g. 2007); Favaro et al. (e.g. 2009). However, for the likelihood ratio assessment it is required not only the probability of observing a new species but also the probability of observing this same species twice (according to the defense the crime stain profile and the suspect profile are two independent observations). Cereda (2015b) is the first paper that addresses the problem of likelihood ratio assessment in the rare haplotype case using Bayesian nonparametric models.

The present paper analyses two frequentist methods, the discrete Laplace method, and a generalization of the Good-Turing, making explicit the corresponding definitions of D, E, and B, and providing a study on the different levels of uncertainty arising for each.

6 The discrete Laplace Method

A discrete random variable X is said to follow the discrete Laplace distribution $\mathrm{DL}(p, y)$, with dispersion parameter $p \in (0, 1)$, and location parameter $y \in \mathbb{Z}$, if its probability density is defined as

$$f(x \mid p, y) = \left(\frac{1-p}{1+p}\right) p^{|x-y|}, \quad \forall x \in \mathbb{Z}.$$

This is used in Andersen et al. (2013b) to model the distribution of single locus Y-STR haplotype in some subpopulation, which is thus assumed to be distributed around a modal allele (represented by the location parameter y).

Each haplotype is actually composed by r loci. Let denote with $\mathbf{X} = (X_1, X_2, ..., X_r)$ the random variable which describes an r-loci haplotype configuration. Moreover, there may be c different subpopulations to take into consideration. By making the strong assumption of independence between loci, within the same subpopulation, the following density is used to describe the probability that $\mathbf{X} = \mathbf{x}$:

$$f(\mathbf{x} \mid \{\mathbf{y_j}\}_j, \{\mathbf{p_j}\}_j) = \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(x_k \mid y_{jk}, p_{jk}),$$

where, for each j, τ_j is the probability a priori of generating from the jth subpopulation, while $\mathbf{p_j} = (p_{j1}, p_{j2}, ..., p_{jr})$ and $\mathbf{y_j} = (y_{j1}, y_{j2}, ..., y_{jr})$ represent the dispersion and location

parameters, respectively, of the jth subpopulation. Andersen et al. (2013b) propose to estimate all these parameters by using using the EM algorithm (Dempster et al., 1977). The initial subpopulation centres are chosen by PAM algorithm (Kaufman and Rousseeuw, 2009) and the number of them by the Bayesian Information Criteria (BIC) (Schwarz, 1978).

6.1 The choice of D in the discrete Laplace Method

The choice of D which underlies the discrete Laplace method, when used to address the rare haplotype match problem is:

- D_{DL} = the particular haplotype x of the suspect and of the stain, along with a database which is a sample from the population of possible perpetrators.

This method allows one to evaluate the data in the light of the usual hypotheses of interest in the DNA example (see Section 2). D_{DL} can be split into E_{DL} and B_{DL}, in the following way:

- E_{DL} = the particular haplotype x of the stain (E_t) and of the suspect (E_s).
- B_{DL} = random sample from the population of possible perpetrators (i.e. database).

The vector containing the frequencies of all haplotypes in the population of reference can be thought of as the nuisance parameter θ of this model. Conditions **a.** and **b.** presented in Section 2 are valid for E_{DL}, B_{DL}, θ, and h, thus the following likelihood ratio (where θ is again omitted) corresponds to this choice of data, evidence, background, and model:

$$\begin{aligned} \mathcal{L}r_{\text{DL}} &= \frac{\mathcal{P}r(D_{\text{DL}} = d \mid h_p)}{\mathcal{P}r(D_{\text{DL}} = d \mid h_d)} = \frac{\mathcal{P}r(E_t = x \mid E_s = x, h_p)\mathcal{P}r(E_s = x \mid h_p)}{\mathcal{P}r(E_t = x \mid E_s = x, h_d)\mathcal{P}r(E_s = x \mid h_d)} \\ &= \frac{\mathcal{P}r(E_t = x \mid E_s = x, h_p)}{\mathcal{P}r(E_t = x \mid h_d)} = \frac{1}{f_x}. \end{aligned} \tag{6}$$

Here, f_x is the frequency of the haplotype x in the population of reference. The second equality is due to conditions **a** and **b** discussed in Section 2.2, while the fourth one is justified by the fact that the distribution of the haplotype of the suspect does not depend on which hypothesis is correct, and that, when θ is fixed (as in the frequentist approach which we are considering) and under h_d, E_t is independent of E_s. The weight of evidence is thus

$$\log_{10} \mathcal{L}r_{\text{DL}} = \log_{10} \frac{1}{f_x}. \tag{7}$$

The frequency f_x can be estimated by $\widehat{f_x}$, using the discrete Laplace method. This brings to the following plug-in estimator for $\log_{10} \mathcal{L}r_{\text{DL}}$:

$$\widehat{\log_{10} \mathcal{L}r_{\text{DL}}} = \log_{10} \frac{1}{\widehat{f_x}}.$$

Notice that the discrete Laplace method uses the database to estimate the number of subpopulations and all the parameters in the model, and this is where B_{DL} comes into play again.

6.2 Quantifying the uncertainty of the discrete Laplace method

We quantify the uncertainty of this method comparing the distribution of $\widehat{\log_{10} \mathcal{L}r_{\mathrm{DL}}} = \log_{10} \frac{1}{\widehat{f_x}}$ with the distribution of the "true" $\log_{10} \mathcal{L}r_{\mathrm{DL}} = \log_{10} \frac{1}{f_x}$. f_x is not known, but we have a database of approximately 19,000 Y-STR 23-loci profiles from 129 different locations in 51 countries in Europe (Purps et al., 2014)[1], which we can pretend contains the whole population of interest for our case. We will consider only 7 loci out of 23 and perform the following experiment: we sample a small database of size $N = 100$, along with a new haplotype (not observed in the small database), and calculate the estimate $\log_{10} \frac{1}{\widehat{f_x}}$. Then, we can use the relative frequency of the haplotype x in the big database as the true one, f_x to obtain $\log_{10} \frac{1}{f_x}$.

This process can be repeated many times (for instance $M = 1000$ samplings of small databases of size $N = 100$ and, for each, a never observed haplotype).

In estimating $\log_{10} \mathcal{L}r_{\mathrm{DL}}$ via $\widehat{f_x}$, one has the choice between adding the haplotype x to the small database before estimating parameters of the discrete Laplace distribution, or not. In a full Bayesian approach the right thing to do is to add the profile to the database. This is shown in Cereda (2015), and we believe that it is the good thing to do also in a frequentist framework. In fact, experiments show that to add or not the haplotype to the database does not make much difference.

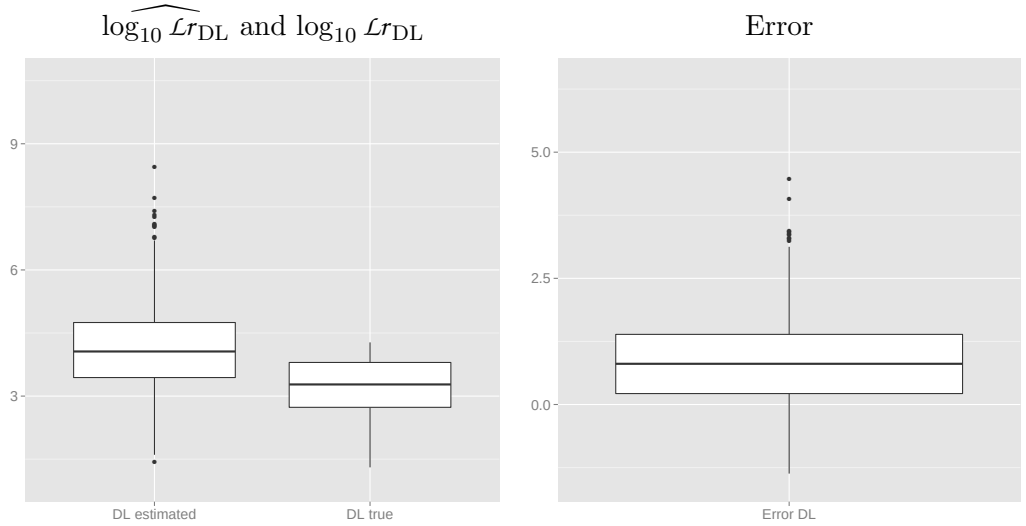
Table 1 and Figure 2 (left part) compare the distributions of $\log_{10} \mathcal{L}r$ and $\widehat{\log_{10} \mathcal{L}r_{\mathrm{DL}}}$, using 7 loci. The same experiment has been carried out for 10 and 3 loci, but not reported in details.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	s.d.
$\log_{10} \mathcal{L}r_{\mathrm{DL}}$	1.305	2.733	3.277	3.272	3.800	4.277	0.666
$\widehat{\log_{10} \mathcal{L}r_{\mathrm{DL}}}$	1.432	3.441	4.061	4.114	4.750	8.452	1.017
Error e_{DL}	-1.37	0.217	0.807	0.842	1.39	4.476	0.863

Table 1: Summaries of the distribution of $\log_{10} \mathcal{L}r_{\mathrm{DL}}$, $\widehat{\log_{10} \mathcal{L}r_{\mathrm{DL}}}$, and of the error e_{DL}.

The error of the discrete Laplace method can be defined as $e_{\mathrm{DL}} := \widehat{\log_{10} \mathcal{L}r_{\mathrm{DL}}} - \log_{10} \mathcal{L}r_{\mathrm{DL}}$. It measures how much the estimated distribution differs from the true one. Table 1 and Figure 2 (right part) show the distribution of the error. One can see that it can attain up to about 4 orders of magnitude. The distribution of the error is mostly lo-

[1] A clean version of the database is provided by Mikkel Meyer Andersen (http://people.math.aau.dk/~mikl/?p=y23).



(a) (b)

Figure 2: discrete Laplace method. Boxplots comparing the distributions of $\log_{10} \mathcal{L}r_{\mathrm{DL}}$ and $\widehat{\log_{10} \mathcal{L}r_{\mathrm{DL}}}$ (left) and the error $e_{\mathrm{DL}} = \widehat{\log_{10} \mathcal{L}r_{\mathrm{DL}}} - \log_{10} \mathcal{L}r_{\mathrm{DL}}$ (2nd column).

cated on positive values, which means that, more often than not, $\widehat{\log_{10} \mathcal{L}r_{\mathrm{DL}}}$ overestimates $\log_{10} \mathcal{L}r_{\mathrm{DL}}$. The standard deviation of the error is small, thereby e_{DL} does not move too much away from its mean, which is about 0.842.

Motivated by the discussion of Section 4, we now analyze the different levels of uncertainty which affect the error. The second level of uncertainty is introduced when the discrete Laplace model, along with all its set of assumptions, is chosen to model the distribution of single locus haplotypes, which in reality do not follow a discrete Laplace distribution.

The third level of uncertainty pertains to the estimation of the parameters of the model (c, p, y, τ). Here, the databases used to estimate the parameters of the discrete Laplace model are probably too small ($N = 100$) with regard to 7 loci.

To decrease both sources of error, one can reduce the number of analyzed loci to 3. The population becomes less sparse, and the databases big enough. Indeed, we performed this experiment and the error decreased a great deal. However, the basic level of uncertainty (see Section 4) is increased inasmuch the data becomes less effective to discern between the two hypotheses. On the other hand, the same experiment with 10 loci leads to obtain more powerful likelihood ratios, but less precise.

The second level of uncertainty can be made harmless assuming an infinite number of subpopulations, since in this way the model will perfectly fit any population. However, this solution will increase the number of parameters, along with the third level of uncertainty.

It is worth underlining that the results of our simulations do not mean that the discrete Laplace method is wrong on the whole, but they show that a blind use of this method is

dangerous. We are applying this method to the specific case of the rare haplotype match, using a database of size 100, and a rather sparse population: maybe this method was never intended to be used for such small databases, and maybe it can be modified in more clever ways to that purpose.

7 The Generalized-Good method

Based on Good (1953), we now propose a nonparametric estimator for the weight of evidence. This is a very good example of data reduction, since $\mathcal{D}$ is here reduced to a greater extent than it was done for the discrete Laplace method. Indeed, the specific haplotype x of the crime stain and of the suspect is ignored, retaining only the fact that they match and the fact that this profile has not been observed yet in the database.

Stated otherwise,

- D_{GG} = the haplotype of the suspect matches the haplotype of the crime stain and it is not in the database.

Consider the following mathematical description: the database of size N can be seen as an i.i.d. sample $(Y_1, Y_2, ..., Y_N)$ from species $\{1, 2, ..., S\}$, with probabilities $(p_1, p_2, ...p_S)$. Hence, the suspect's profile can be though of as the $N+1$st i.i.d. observation. The crime stain's profile is the $N+2$nd observation. According to the defense it is again an i.i.d. draw from $(p_1, p_2, ...p_S)$, while according to prosecution it is equal to the value of Y_{N+1}, with probability one.

The likelihood ratio for this reduction of the data can be thus written as

$$\mathcal{L}r_{\text{GG}} = \frac{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N, Y_{N+1} = Y_{N+2} \mid h_p)}{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N, Y_{N+1} = Y_{N+2} \mid h_d)} = \frac{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N \mid h_p)}{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N, Y_{N+1} = Y_{N+2} \mid h_d)}.$$

From now on, we are presenting results regarding a general database size $N > 2$, and general random variables $Y_1, ..., Y_N$, i.i.d. from $(p_1, p_2, ..., p_S)$. The following notation is used:

$$\theta_1(N; p_1, p_2, ..., p_S) := \mathcal{P}r(Y_N \notin \{Y_1, Y_2, ..., Y_{N-1}\}),$$
$$\theta_2(N; p_1, p_2, ..., p_S) := \mathcal{P}r(Y_N \notin \{Y_1, Y_2, ..., Y_{N-2}\}, Y_N = Y_{N-1}).$$

To make the notation less cumbersome we will use

$$\mathcal{Y}_N := (Y_1, Y_2, ..., Y_N),$$
$$\mathcal{Y}_{i,N} := (Y_1, Y_2, ..., Y_{i-1}, Y_{i+1}, ..., Y_N),$$
$$\mathcal{Y}_{(i,j),N} := (Y_1, Y_2, ..., Y_{i-1}, Y_{i+1}, ..., Y_{j-1}, Y_{j+1}, ..., Y_N), \quad \forall i < j.$$

Moreover, for any random variable Y, and any couple of sets A and B, $\mathbf{1}_{A \cap B^c}(Y)$ is a random variable which has value 1 if Y belongs to the set A and not to the set B, and zero otherwise.

Theorem 1. *An unbiased estimator for* $\theta_1(N; p_1, p_2, ..., p_S)$ *is* $\widehat{\theta}_1(N) = N_1/N$*, where* N_1 *is the number of singletons in the database.*

Proof.

$$\theta_1(N; p_1, p_2, ..., p_S) = \mathcal{P}r(Y_N \notin \mathcal{Y}_{N-1}) = \mathbb{E}(\mathbf{1}_{(\mathcal{Y}_{N-1})^c}(Y_N)) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\mathbf{1}_{(\mathcal{Y}_{i,N})^c}(Y_i))$$
$$= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{(\mathcal{Y}_{i,N})^c}(Y_i)\right) = \mathbb{E}\left(\frac{N_1}{N}\right),$$

where the last equality is due to the fact that the function $\mathbf{1}_{(\mathcal{Y}_{i,N})^c}(Y_i)$ has value 1 for every singleton of the database: the sum is thus the number of singletons (N_1). □

Theorem 2. *An unbiased estimator for* $\theta_2(N; p_1, p_2, ..., p_S)$ *is* $\widehat{\theta}_2(N) = 2N_2/N(N-1)$*, where* N_2 *is the number of doubletons in the database.*

Proof.

$$\theta_2(N; p_1, p_2, ..., p_S) = \mathcal{P}r(Y_N \notin \{\mathcal{Y}_{N-2}\}, Y_N = Y_{N-1}) = \mathbb{E}(\mathbf{1}_{\{Y_{N-1}\cap(\mathcal{Y}_{N-2})^c\}}(Y_N))$$
$$= \frac{2}{N(N-1)}\sum_{i<j}\mathbb{E}(\mathbf{1}_{\{Y_j\cap(\mathcal{Y}_{(i,j),N})^c\}}(Y_i)) = \mathbb{E}\left(\frac{2}{N(N-1)}\sum_{i<j}\mathbf{1}_{\{Y_j\cap(\mathcal{Y}_{(i,j),N})^c\}}(Y_i)\right)$$
$$= \mathbb{E}\left(\frac{2N_2}{N(N-1)}\right),$$

where the last equality is due to the fact that the function $\mathbf{1}_{\{Y_j\cap(\mathcal{Y}_{(i,j),N})c\}}(Y_i)$ has value 1 for each of the N_2 doubletons of the database. □

The two previous theorems can be easily generalized to θ_m defined as $\theta_m(N; p_1, p_2, ..., p_S) := \mathcal{P}r(Y_N \notin \mathcal{Y}_{N-m}, Y_N = Y_{N-1} = .. = Y_{N-m+1})$.

Now we can estimate $\log_{10}\mathcal{L}r_{\mathrm{GG}}$ in the following way:

$$\log_{10}\mathcal{L}r_{\mathrm{GG}} = \log_{10}\frac{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N \mid h_p)}{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N, Y_{N+1} = Y_{N+2} \mid h_d)} \approx \log_{10}\frac{\mathcal{P}r(Y_N \notin \mathcal{Y}_{N-1})}{\mathcal{P}r(Y_N \notin \mathcal{Y}_{N-2}, Y_N = Y_{N-1})}$$
$$\approx \log_{10}\frac{\theta_1(N; p_1, p_2, ..., p_S)}{\theta_2(N; p_1, p_2, ..., p_S)}.$$

Thus, we propose the following estimator for the weight of evidence:

$$\widehat{\log_{10}\mathcal{L}r_{\mathrm{GG}}} = \log_{10}\frac{\widehat{\theta}_1(N)}{\widehat{\theta}_2(N)} = \log_{10}\frac{(N-1)N_1}{2N_2} \approx \log_{10}\frac{NN_1}{2N_2}. \tag{8}$$

Notice that there are two kinds of approximation steps: a mathematical approximation of $\theta_1(N+1; p_1, p_2, ..., p_S)$ with $\theta_1(N; p_1, p_2, ..., p_S)$, which should hardly make any difference, for reasonably large N, and a statistical estimation of $\theta_1(N; p_1, p_2, ..., p_S)$ using an unbiased estimator (and similarly for θ_2).

It is important to underline that, due to Jensen's inequality, the estimators $\log_{10}\widehat{\theta_1}$ and $\log_{10}\widehat{\theta_2}$ are not unbiased for $\log_{10}\theta_1$ and $\log_{10}\theta_2$, but it will be shown by simulations that $\widehat{\log_{10}\mathcal{L}r_{\mathrm{GG}}}$ is approximately unbiased for $\log_{10}\mathcal{L}r_{\mathrm{GG}}$. However, the point is not to find an unbiased estimator, but one with a small error rate.

Notice that in order to estimate $\log_{10}\mathcal{L}r_{\mathrm{GG}}$ it is not necessary to use all the information contained in the database, but only N, N_1, and N_2, that is the number of singletons and doubletons in the database. The nuisance parameter of the model is the vector θ containing the frequencies of the Y-STR haplotypes in the population of interest. θ_1 and θ_2 are functions of θ.

The limitation of this method is that it cannot be used if $N_2 = 0$ (this corresponds to an infinite likelihood ratio) and it does not perform well also in case the number of singletons is very small or zero. We believe it can be improved and extended by smoothing techniques (Good, 1953; Anevski et al., 2013), but we are going to ignore this problem.

The 'κ-method' of Brenner (Brenner, 2010) is based on an analogous line of reasoning. It estimates the likelihood ratio as $\widehat{\mathcal{L}r}_\kappa = \frac{N^2}{N-N_1}$. However, in the derivation of this estimator, there is an approximation involved, based on assumptions which are not always satisfied, leading sometimes to anti-conservatism (see also the discussion in Buckleton et al. (2011), and the answer in Brenner (2014)). In particular, Brenner (2014) provides a pathological population where the approximation does not hold, while showing empirical evidence that for Fisher-Wright populations the condition is fulfilled. Our method is, on the other hand, based on a principled derivation of the estimator of equation (8), which is similar to Brenner's one under the following conditions: there are almost only singletons and doubletons in the database, and $N_1 \gg N_2$. These assumptions are typically satisfied, explaining why Brenner's method often works. They also constitute a good description of when it does not work.

Lastly, we remark that this method can be generalized in the obvious way, to the case in which the haplotype is indeed in the database. Moreover, this method is suitable to be directly applied to different kinds of evidence.

7.1 Quantifying the uncertainty of the GG method

As we did in Section 6.2, we want to quantify the uncertainty of this method. One way is to compare the distribution of

$$\widehat{\log_{10}\mathcal{L}r_{\mathrm{GG}}} = \log_{10}\frac{NN_1}{2N_2},$$

with the distribution of the "true"

$$\log_{10}\mathcal{L}r_{\mathrm{GG}} = \log_{10}\frac{\mathcal{P}r(Y_{N+1}\notin\mathcal{Y}_N)}{\mathcal{P}r(Y_{N+1}\notin\mathcal{Y}_N\cap Y_{N+1}=Y_{N+2})} := \log_{10}\frac{\theta_1}{\theta_2}.$$

Actually, the latter is not a distribution, but a single value, unknown. Again, we pretend that the database of Purps et al. (2014) contains the profiles of the whole population, to find out the 'true' θ_1 and θ_2, restricting our simulations to 7 loci. To do so, we sample M small databases of size $N = 100$, along with two other haplotypes. θ_1 is the proportion of times in which the $(N+1)$st haplotype is a new one (i.e., not one of the previous N), and θ_2 is the proportion of times in which the $(N+2)$nd is equal to the $(N+1)$st, and different from the first N observations. From our simulations, we used $M = 100,000$, and we obtained θ_1, θ_2, and $\log_{10} \mathcal{L}r$ as in Table 2.

θ_1	θ_2	True $\log_{10} \mathcal{L}r_{\text{GG}}$
0.748	0.0012	2.78

Table 2: Values of θ_1 and θ_2 and of $\log_{10} \mathcal{L}r_{\text{GG}}$ obtained by simulations, assuming that the database of Purps et al. (2014) contains the whole population of interest.

The distribution of $\widehat{\log_{10} \mathcal{L}r_{\text{GG}}} = \log_{10} \frac{NN_1}{2N_2}$ can be obtained by sampling $M = 100,000$ databases of size $N = 100$. Out of 100,000 databases, 121 had $N_2 = 0$. They have been removed from the data, and we acknowledge that this choice creates unfairness to the discrete Laplace method. On the other hand, we believe that this occurs frequently enough not to affect very strongly the comparison.

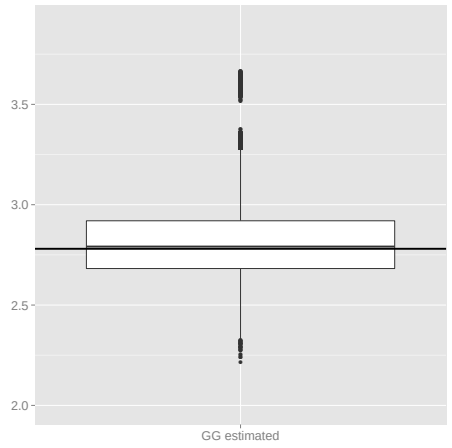


Figure 3: Boxplots of the distribution of $\widehat{\log_{10} \mathcal{L}r_{\text{GG}}}$ around the true value $\log_{10} \mathcal{L}r_{\text{GG}}$ (black line).

Figure 3 shows the distribution of the estimator $\widehat{\log_{10} \mathcal{L}r_{\text{GG}}}$ around the true value (black line). The error of the Generalized-Good method, defined as $e_{\text{GG}} = \widehat{\log_{10} \mathcal{L}r_{\text{GG}}} - \log_{10} \mathcal{L}r_{\text{GG}}$, tells us how much the estimator differs from the true value.

Table 3 provides the summaries for $\widehat{\log_{10} \mathcal{L}r_{\text{GG}}}$, and for the error e_{DL}. We don't provide the plots for the distribution of e_{GG} since they are identical to those in Figure 3, shifted

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd
$\log_{10} \mathcal{L}r_{\mathrm{GG}}$	2.78	2.78	2.78	2.78	2.78	2.78	0
$\widehat{\log_{10} \mathcal{L}r_{\mathrm{GG}}}$	2.215	2.682	2.792	2.818	2.920	3.668	0.198
Error e_{GG}	-0.566	-0.098	0.0112	0.038	0.14	0.887	0.198

Table 3: Summaries of the distribution of $\widehat{\log_{10} \mathcal{L}r_{\mathrm{GG}}}$, of $\log_{10} \mathcal{L}r_{\mathrm{GG}}$, and of the error e_{GG}.

of $\log_{10} \mathcal{L}r_{\mathrm{GG}}$.

One can see that the error can attain up to about 0.9 orders of magnitude. The distribution of the error is mostly located on the positive values, which means that, more often than not, $\widehat{\log_{10} \mathcal{L}r_{\mathrm{GG}}}$ overestimates $\log_{10} \mathcal{L}r_{\mathrm{GG}}$. The standard deviation of the error is small, thereby e_{GG} does not move too much away from the mean, which is about 0.038. If compared to the error of the discrete Laplace method, one can conclude that here we get a better estimator in terms of accuracy, since the error ranges over more restrained values and the standard deviation is much smaller. However, it is important to keep in mind that they are not different estimators of the same quantity, but different estimators of different quantities, since the reduction of data used by the Generalized-Good method, which allows obtaining accuracy in the estimates is less strong to discern between the two hypotheses.

8 Choosing and comparing methods

In comparing the two methods one can consider the precision with respect to what the method is trying to estimate, quantified by the errors e_{DL}, and e_{GG}. These errors are due to the two second and third level of uncertainty described in Section 4, and decrease sensibly if data is reduced. This is why, under this aspect, the Generalized-Good is to be preferred to the discrete Laplace, and for the latter a fewer number of loci is to be preferred. However, it is not correct to believe that the greater the reduction, the better is the method. To reduce means to lose information, and thus to diminish the capability of the method to distinguish between the hypotheses at stake (the first, or basic level of uncertainty). In order to investigate this loss, one can compare each method to the likelihood ratio $1/f$ (where f is the population frequency of the matching haplotype), which can be considered the hallmark in a population with no substructure. Comparing Table 1 with Table 2 one can see for instance that choosing the Generalized-Good one loses on average around 0.5 (in logarithmic scale) in terms of strength of data to discriminate between hypotheses. This is a small disadvantage for the prosecution, while everybody gain in terms of precision with respect to the true $\log_{10} \mathcal{L}r_{\mathrm{GG}}$. As a last remark, we invite the reader to realize that the discrete Laplace method is better inasmuch it can always be

used. On the other hand, for the Generalized-Good, we had to remove 121 experiments where $N_2 = 0$.

9 Remark and conclusion

The aim of this paper could, at first sight, be considered that of offering two additional frequentist methods to address the issue of the likelihood ratio calculation in the case of a rare Y-STR haplotype match. However, a careful reader may have realized that these methods also constitute two interesting opportunities to show and apply the guidelines exposed in the opening sections. In particular, two important facts are pointed out in Sections 3 and 4: first, it is more sensible to talk about "a" likelihood ratio instead of "the" likelihood ratio, and second, a quantification of the error involved in the estimation is to be provided along with the estimate of the likelihood ratio.

Moreover, it is explained that sometimes it is possible to the break down the data to be evaluated into E (which is sufficient for H) and B (which is irrelevant for H). The discrete Laplace method (developed in Section 6) is a good example where this distinction can be done, while the same is not true for the Generalized-Good method (Section 7).

Lastly, this paper wants to get across the message that reducing the data to a smaller extent is sometimes not only necessary, but also desirable in terms of exactitude of the estimates, as proved by the comparison between the discrete Laplace method (less reduction, less precision of the estimates) and the Generalized-Good method (stronger reduction, more precision of the estimates). In this respect we disagree with Buckleton et al. (2011) who, talking about Brenner's method, state that 'there is a merit focussing in the type or name of a lineage marker". Although we agree that "such ignorance of type implies a substantial loss of information", it may allow a large gain in precision.

The take home message is that choosing the best method is clearly a very delicate task. One has to consider many different aspects, and look for a compromise which is acceptable for the specific application at hand. It is important to realise that in this paper we study a very extreme situation with very small databases and a possibly unrealistic population, for which the Generalized-Good seemed to be the best compromise. Clearly, there are no possible general conclusions to be given, other than at each new situation one has to reconsider all these aspects, and weigh them.

Acknowledgements

The Generalized-Good method described here was suggested by Richard Gill and presented in several conference lectures, see for instance http://www.slideshare.net/gill1109/the-fundamental-problem-of-forensic-statistics-38322519. I am indebted to Charles

Brenner (the first to use the 'fundamental problem' name), and to Ronald Meester, for the useful discussions about this paper, which lead to many improvements. This research was supported by the Swiss National Science Foundation, through grants no. 105311-1445570 and 10531A-156146/1, and carried out in the context of a joint research project, supervised by Franco Taroni (University of Lausanne, Ecole des sciences criminelles, Faculté de droit, des sciences criminelles et d'administration publique), and Richard Gill (Mathematical Institute, Leiden University).

References

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

Aitken, C. G. G., Taroni, F., Barnett, P. D., and Tsatsakis, A. M. (1998). A verbal scale for the interpretation of evidence. *Science & Justice*, 38:279–283.

Andersen, M. M., Caliebe, A., Jochens, A., Willuweit, S., and Krawczak, M. (2013a). Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics*, 7:264–271.

Andersen, M. M., Eriksen, P. S., and Morling, N. (2013b). The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, 329:39–51.

Anevski, D., Gill, R. D., and Zohren, S. (2013). Estimating a probability mass function with unknown labels. http://arxiv.org/abs/1312.1200.

Balding, D. (2005). *Weight-of-evidence for Forensic DNA Profiles*. John Wiley & Sons Hoboken, NJ.

Brenner, C. H. (2010). Fundamental problem of forensic mathematics—The evidential value of a rare haplotype. *Forensic Science International: Genetics*, 4:281–291.

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

Buckleton, J., Krawczak, M., and Weir, B. (2011). The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics*, 5:78–83.

Cereda, G. (2015). Bayesian approach to LR assessment in case of rare type match: careful derivation and limits. arXiv:1502.02406.

Cereda, G. (2015b). Nonparametric Bayesian approach to LR assessment in case of rare haplotype match. arxiv:1506.08444.

Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014). An investigation of the potential of DIP-STR markers for DNA mixture analyses. *Forensic Science International: Genetics*, 11:229 – 240.

Curran, J., Buckleton, J., Triggs, C., and Weir, B. (2002). Assessing uncertainty in DNA evidence caused by sampling effects. *Science & Justice*, 42:29–37.

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

Dawid, P. (2001). Comment on Stockmarr's "Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search" *Biometrics*, 57:976–980.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–38.

Egeland, T. and Salas, A. (2008). Estimating haplotype frequency and coverage of databases. *PLoS ONE*, 3:e3988–e3988.

Evett, I. and Weir, B. (1998). *Interpreting DNA evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates, Sunderland.

Favaro, S., Lijoi, A., Mena, R. H., and Pruenster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Methodological)*, 71:993–1008.

Gale, W. A. and Church, K. W. (1994). What's wrong with adding one? In *Corpus-Based Research into Language. Rodolpi*.

Gill, P., Gusmão, L., Haned, H., Mayr, W. R., Morling, N., Parson, W., Prieto, L., Prinz, M., Schneider, H., Schneider, P. M., and Weir, B. S. (2012). DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Science International: Genetics*, 6:679–688.

Good, I. (1950). *Probability and the Weighing of Evidence*. Charles Griffin, London.

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264.

Kaufman, L. and Rousseeuw, P. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley.

Krawczak, M. (2001). Forensic evaluation of Y-STR haplotype matches: a comment. *Forensic Science International*, 118:114–115.

Krichevsky, R. and Trofimov, V. (1981). The performance of universal coding. *IEEE Transactions on Information Theory*, 27:199–207.

Laplace, P. (1814). *Essai philosophique sur les probabilités*. Mme Ve. Courcier.

Lijoi, A., Mena, R. H., and Pruenster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *BIOMETRIKA*, 94:769–786.

Louis, T. A. (1981). Confidence intervals for a binomial parameter after observing no successes. *The American Statistician*, 35:154–154.

Morrison, G. S. (2010). *Evidence Expert*, chapter Forensic voice comparison. Thomson Reuters, Sidney, Australia.

Orlitsky, A., Santhanam, N., and Zhang, J. (2003). Always Good Turing: asymptotically optimal probability estimation. *Science (New York, N.Y.)*, 302:427–431.

Orlitsky, A., Santhanam, N. P., Viswanathan, K., and Zhang, J. (2004). On Modeling Profiles Instead of Values. In *UAI*, pages 426–435.

Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Angustia, S. M. T., Santos, L. H., Anslinger, K., Bayer, B., Ayub, Q., Wei, W., Xue, Y., Tyler-Smith, C., Bafalluy, M. B., Martínez-Jarreta, B., Egyed, B., Balitzki, B., Tschumi, S., Ballard, D., Court, D. S., Barrantes, X., Bäßler, G., Wiest, T., Berger, B., Niederstätter, H., Parson, W., Davis, C., Budowle, B., Burri, H., Borer, U., Koller, C., Carvalho, E. F., Domingues, P. M., Chamoun, W. T., Coble, M. D., Hill, C. R., Corach, D., Caputo, M., D'Amato, M. E., Davison, S., Decorte, R., Larmuseau, M. H. D., Ottoni, C., Rickards, O., Lu, D., Jiang, C., Dobosz, T., Jonkisz, A., Frank, W. E., Furac, I., Gehrig, C., Castella, V., Grskovic, B., Haas, C., Wobst, J., Hadzic, G., Drobnic, K., Honda, K., Hou, Y., Zhou, D., Li, Y., Hu, S., Chen, S., Immel, U.-D., Lessig, R., Jakovski, Z., Ilievska, T., Klann, A. E., García, C. C., de Knijff, P., Kraaijenbrink, T., Kondili, A., Miniati, P., Vouropoulou, M., Kovacevic, L., Marjanovic, D., Lindner, I., Mansour, I., Al-Azem, M., Andari, A. E., Marino, M., Furfuro, S., Locarno, L., Martín, P., Luque, G. M., Alonso, A., Miranda, L. S., Moreira, H., Mizuno, N., Iwashima, Y., Neto, R. S. M., Nogueira, T. L. S., Silva, R., Nastainczyk-Wulf, M., Edelmann, J., Kohl, M., Nie, S., Wang, X., Cheng, B., Núñez, C., Pancorbo, M. M. d., Olofsson, J. K., Morling, N., Onofri, V., Tagliabracci, A., Pamjav, H., Volgyi, A., Barany, G., Pawlowski, R., Maciejewska, A., Pelotti, S., Pepinski, W., Abreu-Glowacka, M., Phillips, C., Cárdenas, J., Rey-Gonzalez, D., Salas, A., Brisighelli, F., Capelli, C., Toscanini, U., Piccinini, A., Piglionica, M.,

Baldassarra, S. L., Ploski, R., Konarzewska, M., Jastrzebska, E., Robino, C., Sajantila, A., Palo, J. U., Guevara, E., Salvador, J., Ungria, M. C. D., Rodriguez, J. J. R., Schmidt, U., Schlauderer, N., Saukko, P., Schneider, P. M., Sirker, M., Shin, K.-J., Oh, Y. N., Skitsa, I., Ampati, A., Smith, T.-G., Calvit, L. S. d., Stenzl, V., Capal, T., Tillmar, A., Nilsson, H., Turrina, S., De Leo, D., Verzeletti, A., Cortellini, V., Wetton, J. H., Gwynne, G. M., Jobling, M. A., Whittle, M. R., Sumita, D. R., Wolańska-Nowak, P., Yong, R. Y. Y., Krawczak, M., Nothnagel, M., and Roewer, L. (2014). A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Science International: Genetics*, 12:12–23.

Robertson, B. and Vignaux, G. A. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. John Wiley & Sons, Chichester.

Roewer, L., Kayser, M., de Knijff, P., Anslinger, K., Betz, A., Caglia, A., Corach, D., Furedi, S., Henke, L., Hidding, M., Kargel, H., Lessig, R., Nagy, M., Pascali, V., Parson, W., Rolf, B., Schmitt, C., Szibor, R., Teifel-Greding, J., and Krawczak, M. (2000). A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International*, 114:31–43.

Schum, D. (1994). *The Evidential Foundations of Probabilistic Reasoning*. Northwestern University Press, Evanston.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Sjerps, M. J., Alberink, I., Bolck, A., Stoel, R. D., Vergeer, P., and van Zanten, J. H. (2016). Uncertainty and LR: to integrate or not to integrate, that's the question. *Law, Probability and Risk*, 15:23–29.

Steele, C. D. and Balding, D. J. (2014). Statistical evaluation of forensic DNA profile evidence. *Annual Review of Statistics and Its Application*, 1:361–384.

Stoel, R. D. and Sjerps, M. (2012). Interpretation of forensic evidence. In *Handbook of Risk Theory*, pages 135–158. Springer Netherlands.

Taroni, F., Biedermann, A., Bozza, S., Garbolino, P., and Aitken, C. (2014). *Bayesian networks for probabilistic inference and decision analysis in forensic science*. John Wiley & Sons, Chichester, second edition.

Taroni, F., Bozza, S., Biedermann, A., and Aitken, C. (2015). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, 0:1–16.

Taroni, F., Bozza, S., Biedermann, A., Garbolino, P., and Aitken, C. (2010). *Data Analysis in Forensic Science: A Bayesian Decision Perspective*. Statistics in Practice. Wiley, Chichester.

Tiwari, R. C. and Tripathi, R. C. (1989). Nonparametric Bayes estimation of the probability of discovering a new species. *Communications in statistics: Theory and methods*, A18:877–895.

Willuweit, S., Caliebe, A., Andersen, M. M., and Roewer, L. (2011). Y-STR Frequency Surveying Method: A critical reappraisal. *Forensic Science International: Genetics*, 5:84–90.