Forensic Science International: Genetics 14 (2015) 116–124

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Efficient computations with the likelihood ratio distribution

Maarten Kruijver *

Department of Mathematics, VU University, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 12 February 2014
Received in revised form 4 August 2014
Accepted 23 September 2014

Keywords:
Likelihood ratio
Exceedance probability
Exact computation
Monte Carlo
Simulation
Importance sampling

ABSTRACT

What is the probability that the likelihood ratio exceeds a threshold *t*, if a specified hypothesis is true? This question is asked, for instance, when performing power calculations for kinship testing, when computing true and false positive rates for familial searching and when computing the power of discrimination of a complex mixture. Answering this question is not straightforward, since there is are a huge number of possible genotypic combinations to consider. Different solutions are found in the literature. Several authors estimate the threshold exceedance probability using simulation. Corradi and Ricciardi [1] propose a discrete approximation to the likelihood ratio distribution which yields a lower and upper bound on the probability. Nothnagel et al. [2] use the normal distribution as an approximation to the likelihood ratio distribution. Dørum et al. [3] introduce an algorithm that can be used for exact computation, but this algorithm is computationally intensive, unless the threshold *t* is very large.

We present three new approaches to the problem. Firstly, we show how importance sampling can be used to make the simulation approach significantly more efficient. Importance sampling is a statistical technique that turns out to work well in the current context. Secondly, we present a novel algorithm for computing exceedance probabilities. The algorithm is exact, fast and can handle relatively large problems. Thirdly, we introduce an approach that combines the novel algorithm with the discrete approximation of Corradi and Ricciardi. This last approach can be applied to very large problems and yields a lower and upper bound on the exceedance probability. The use of the different approaches is illustrated with examples from forensic genetics, such as kinship testing, familial searching and mixture interpretation. The algorithms are implemented in an R-package called *DNAprofiles*, which is freely available from CRAN.

1. Introduction

Likelihood ratios [4] are used to assess the evidential value of a piece of evidence under competing hypotheses. Applications of forensic genetics that use the likelihood ratio include identification of persons [5], relationship testing [6], familial database searching [7] and DNA mixture interpretation [8]. The question addressed here is: what is the probability of obtaining a likelihood ratio exceeding a given threshold, if a specific hypothesis is true? This question arises in several applications. For instance, in the context of DNA mixtures: what is the probability that a non-contributor obtains a likelihood ratio exceeding 1000 [9]? Or, considering paternity testing, many labs have the policy of deciding on paternity when the likelihood ratio exceeds 10,000 [10]. One might ask: what is the probability for a true father to have a likelihood ratio exceeding this threshold (the true positive rate)? And what is the same probability for an unrelated person (the false positive rate)? Many other practical questions ask for the probability of exceeding a likelihood ratio threshold under a given hypothesis.

Computing threshold exceedance probabilities for likelihood ratios is not straightforward and requires significant computational effort. The literature contains several approaches. One approach is to use simulation [11–13] to estimate exceedance probabilities. Specifically, several authors simulate a large number of genotypes according to a hypothesis and estimate the threshold exceedance probability from the empirical fraction seen in the random samples. Simulation is relatively easily implemented, but a drawback is that many simulations are needed to estimate small probabilities reliably. A second possibility is to approximate the likelihood ratio distribution with a continuous distribution. Nothnagel et al. [2] approximate the distribution of the logarithm of the likelihood ratio with the normal distribution. The normal approximation can be very poor in terms of exceedance probabilities, especially for large values of the threshold. A third possibility is to approximate the likelihood ratio distribution with a discrete distribution that has fewer outcomes than the original

* Tel.: +31 633850120.
E-mail address: m.v.kruijver@vu.nl

http://dx.doi.org/10.1016/j.fsigen.2014.09.018
1872-4973/

distribution. This is the distribution that Corradi and Ricciardi use [1], who provide a method to compute lower and upper bounds on the exceedance probabilities. This paper shows how their method can be improved and that it can be extended further when combined with the novel algorithm for exact computation that is presented here. A fourth approach is presented by Dørum et al. [3], who propose an algorithm that computes exceedance probabilities and requires very few computations for very large values of the threshold. Their algorithm however can be prohibitively computationally expensive unless the threshold value is very large.

This paper introduces three new approaches to the problem of computing exceedance probabilities. Firstly, the paper explains how to speed up the simulation approach by applying importance sampling. This statistical technique [14] makes it feasible to estimate small probabilities reliably using simulation. Dørum et al. [3] write that importance sampling appears to be difficult in the forensic genetics context. As demonstrated here, however, the application of importance sampling turns out to be straightforward. Secondly, a new algorithm is presented that can be used to compute exceedance probabilities. The algorithm divides the loci into two subsets and explicitly calculates the probability distribution of the combined likelihood ratio at the two subsets. Exceedance probabilities for the combined likelihood ratio on all the loci can be computed efficiently from the probability distributions at the two subsets. The algorithm can be applied as long as the probability distributions at the two subsets can be stored. Thus, the algorithm can be applied to larger problems than the algorithm of Dørum et al. but the largest problems are still out of reach. Therefore, thirdly, a new algorithm is proposed that iteratively applies the approximation of Corradi and Ricciardi [1], until exceedance probabilities can be computed with the newly introduced exact algorithm. As such, the algorithm can be considered a refinement of Corradi and Ricciardi's approach. Like the latter, it can be applied to the largest problems and yields a lower and upper bound for the exceedance probability. The three new approaches are illustrated using various examples from forensic genetics.

The structure of this paper is as follows. Section 2 introduces the problem of determining exceedance probabilities for likelihood ratios, after which Section 3 discusses approaches from the literature. Section 4 explains how importance sampling can be used to speed up the simulation approach. The new algorithm for computing exceedance probabilities is presented in Section 5. The new algorithm that iteratively approximates the likelihood ratio distribution is introduced in Section 6. Section 7 applies the three new approaches to several problems from forensic genetics. The conclusion follows in Section 8.

2. Problem description

The likelihood ratio is the ratio of the probability of the evidence under the prosecution hypothesis (H_p) and the defense hypothesis (H_d) [15]. The likelihood ratio at locus i is given by:

$$\mathrm{LR}^{(i)}(e_i) = \frac{\mathrm{P}(e_i|H_p)}{\mathrm{P}(e_i|H_d)}, \tag{2.1}$$

where e_i denotes the evidence at locus i. For instance, if a single-contributor DNA profile is observed at m loci with codominant alleles, then $e_i = (a, b)$, with a and b being the alleles of the profile at locus i. By convention we call the two hypotheses the prosecution and the defense hypothesis, even outside the criminal context. For independent loci, the combined likelihood ratio is computed as the product of the likelihood ratios at each locus:

$$\mathrm{LR}(e) = \frac{\mathrm{P}(e|H_p)}{\mathrm{P}(e|H_d)} = \prod_{i=1}^{m} \frac{\mathrm{P}(e_i|H_p)}{\mathrm{P}(e_i|H_d)} = \prod_{i=1}^{m} \mathrm{LR}^{(i)}(e_i). \tag{2.2}$$

The problem is how to compute the probability that the combined likelihood ratio exceeds a threshold t, if an hypothesis H is true. This probability is denoted:

$$q_{t|H} := \mathrm{P}(\mathrm{LR} > t|H), \tag{2.3}$$

with the dependence on e omitted for brevity. Conversely, the probability that the combined likelihood ratio equals at most t is denoted:

$$p_{t|H} := \mathrm{P}(\mathrm{LR} \leq t|H) = 1 - q_{t|H}. \tag{2.4}$$

The hypothesis H can be H_p, H_d or even another hypothesis. To avoid division by zero, every genotype with non-zero probability under H_p must also have non-zero probability under H_d. Likewise, we assume that every genotype that can be seen under H_p, can also be seen under H.

The likelihood ratio at each locus can take a small number of possible values, called outcomes in the probability context. For instance, when the evidence is single locus DNA profile, there can be at most $\binom{A+1}{2} = A \times (A+1)/2$ outcomes of the likelihood ratio, where A denotes the length of the allele ladder. There can be less than $\binom{A+1}{2}$ outcomes per locus, since the likelihood ratio can be identical for several genotypes, as will be shown in Section 2.2. Denote by n_i the number of outcomes at locus i. The outcomes are denoted $x_1^{(i)}, \ldots, x_{n_i}^{(i)}$ and, under H, the outcomes occur with probabilities $f_1^{(i)}, \ldots, f_{n_i}^{(i)}$. Formally:

$$\mathrm{P}(\mathrm{LR}^{(i)} = x_j^{(i)}|H) = f_j^{(i)}, \quad \text{for } j = 1, \ldots, n_i. \tag{2.5}$$

For independent loci, the combined likelihood ratio is the product of the m likelihood ratios at each locus, so it can be written as:

$$\mathrm{LR} = \mathrm{LR}^{(1)}\mathrm{LR}^{(2)} \cdots \mathrm{LR}^{(m)}. \tag{2.6}$$

The probability that, under H, the combined likelihood ratio exceeds t can then be written as:

$$\begin{aligned} q_{t|H} &= \mathrm{P}(\mathrm{LR} > t|H) \\ &= \sum_{i_1, \ldots, i_m} f_{i_1}^{(1)} f_{i_2}^{(2)} \cdots f_{i_m}^{(m)} \mathbf{1}\{x_{i_1}^{(1)} x_{i_2}^{(2)} \cdots x_{i_m}^{(m)} > t\}, \end{aligned} \tag{2.7}$$

where $\mathbf{1}$ denotes the indicator function. The sum involves $\prod_{i=1}^{m} n_i$ terms. Hence, direct evaluation is only feasible when a few loci are used or the number of outcomes per locus is small. The following two examples illustrate the number of terms that can be involved in the sum. The first example concerns the distribution of the likelihood ratio for pairwise kinship with a fixed person. The likelihood ratio thus depends on the single DNA profile of the person that is not fixed, which leads to a sum that involves from 10^7 to over 10^{11} terms, depending on whether 10 or 15 loci are used. The second example concerns again the likelihood ratio for pairwise kinship, but does not assume one of the two persons to be fixed. The likelihood ratio now depends on two DNA profiles, which leads to a sum that involves over 10^{20} terms when 10 loci are used and over 10^{30} terms for 15 loci. It is completely unfeasible to compute these sums directly, even on a fast computer.

For the sake of the exposition, the following two examples as well as most other examples use a simple genetic model that disregards linkage, contamination, stutter, dropin, dropout and subpopulation effects. The approaches described here can also be used with more sophisticated models, although this is less straightforward, since computing the likelihood ratio then becomes a problem on its own.