Forensic Science International: Genetics 27 (2017) 74–81

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Importance sampling allows H_d true tests of highly discriminating DNA profiles

Duncan Taylor[a,b,*], James M. Curran[c], John Buckleton[d,e]

[a] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[b] School of Biological Sciences, Flinders University, GPO Box 2100 Adelaide SA, Australia 5001 ESR, Private Bag 92021, Auckland 1142, New Zealand
[c] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[d] National Institute of Standards and Technology, 100 Bureau Drive, MS 8980 and 8314, Gaithersburg, MD 20899, United States
[e] ESR, Private Bag 92021, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 21 July 2016
Received in revised form 1 December 2016
Accepted 8 December 2016
Available online 9 December 2016

Keywords:
DNA profile
H_d true
Likelihood ratios
Performance tests
Importance sampling

ABSTRACT

H_d true testing is a way of assessing the performance of a model, or DNA profile interpretation system. These tests involve simulating DNA profiles of non-donors to a DNA mixture and calculating a likelihood ratio (*LR*) with one proposition postulating their contribution and the alternative postulating their non-contribution. Following Turing it is possible to predict that "*The average LR for the H_d true tests should be one*" [1]. This suggests a way of validating softwares. During discussions on the ISFG software validation guidelines [2] it was argued by some that this prediction had not been sufficiently examined experimentally to serve as a criterion for validation. More recently a high profile report [3] has emphasised large scale empirical examination.

A limitation with H_d true tests, when non-donor profiles are generated at random (or in accordance with expectation from allele frequencies), is that the number of tests required depends on the discrimination power of the evidence profile. If the H_d true tests are to fully explore the genotype space that yields non-zero *LR*s then the number of simulations required could be in the 10 s of orders of magnitude (well outside practical computing limits). We describe here the use of importance sampling, which allows the simulation of rare events to occur more commonly than they would at random, and then adjusting for this bias at the end of the simulation in order to recover all diagnostic values of interest. Importance sampling, whilst having been employed by others for H_d true tests, is largely unknown in forensic genetics. We take time in this paper to explain how importance sampling works, the advantages of using it and its application to H_d true tests. We conclude by showing that employing an importance sampling scheme brings H_d true testing ability to all profiles, regardless of discrimination power.

1. Introduction

A recent publication [1] examined a method of simulation-based performance testing of a model [4,5] used to evaluate DNA profiling data. These tests involved simulating DNA profiles of non-donors to a DNA mixture and calculating a likelihood ratio (*LR*) with one proposition postulating their contribution and the alternative postulating their non-contribution. Tests simulating the situation where a person of interest (POI) is not a DNA donor are more appropriately called 'H_d true' tests rather than performance tests. Good [6] (quoting Turing) stated "*the expected factor for a wrong hypothesis in virtue of any experiment is 1.*" Focussing this to the problem at hand translates to "*The average LR for the H_d true tests should be one*". In [1] the truth of this lemma was demonstrated by the use of H_d true tests on nine DNA profiles of varying complexity and information content This suggests a way of validating softwares by noting the average *LR* in a large number of Hd true tests.[1] During discussions on the ISFG software validation

[1] Note that adherence to this lemma is not the only test that a system would need to pass in order to be considered valid. The adherence of a system to the lemma follows from the laws of probability, hence while it will demonstrate that a probability distribution has been formed on the genotypes it does not mean that the probability distribution is sensible. Secondly, we do not know how systems will behave that treat nuisance parameters in the model differently under Hp and Hd, but it is quite probable that they will not adhere to the lemma. The implications of this behaviour are beyond the scope of this article.

* Corresponding author at: Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia.
E-mail address: Duncan.Taylor@sa.gov.au (D. Taylor).

http://dx.doi.org/10.1016/j.fsigen.2016.12.004
1872-4973/

guidelines [2] it was argued by some that this prediction had not been sufficiently examined experimentally to serve as a criterion for validation. More recently a high profile report [3] has emphasised large scale empirical examination.

The discrimination power of the DNA profiles that can be tested by standard sampling is limited. An *LR* of *x* requires simulation that has many more than *x* elements. For a complete DNA profile in a modern profiling system the value of *x* can be over 20 orders of magnitude, which is well beyond the practical limits of any standard computer using a naïve simulator.

In the simulations carried out in [1] on profiles with highly discriminating information the vast majority of *LR*s produced had a value of zero. A situation can be imagined where a single source DNA profile that had a profile frequency of 1 in 1 billion was undergoing H_d true tests using propositions:

H_p The randomly simulated non-donor is the source of the DNA

H_d An unknown individual is the source of the DNA

Within each block of one billion tests we would expect an *LR* of one billion to be obtained once, and the rest of the simulations would yield an *LR* of zero. Most observers would agree that this seems like a large effort to obtain mostly zeros. A more efficient system would simulate profiles that we knew were not going to yield an *LR* of zero, and as long as we knew what proportion would give zero (had we carried out the naïve simulation) then we would end up with the same total information as using a naïve simulator. The advantage, however, is a very much reduced requirement for simulation. In the single source example described above, we would only need to run one test that simulated the one profile that gave an inclusionary *LR* and as long as we know that the frequency of obtaining a non-zero *LR* was1×10^{-9}, then we would have the same information as before but at 1 billionth the computing cost. This is the idea behind a technique known as 'importance sampling'.

This idea is not new to forensic biology. In a recent publication describing the workings of continuous DNA interpretation software [7], importance sampling was used to consider genotypes that are included in an assessment of the probability of the evidence. In [8] the authors demonstrate the workings of importance sampling as applied to choosing genotypes to calculate the proportion of *LR*s derived from mixtures above a chosen value. Prior to this importance sampling was very nicely demonstrated in [9] with application to calculating exceedance probabilities. Despite these publications, the idea of importance sampling can be difficult to understand for those who do not have a statistical background. We attempt, in this work, to explain what importance sampling is, with simple examples, and how it is beneficial when using a sampling system to assign a probability for the occurrence of rare events.

We demonstrate the application of importance sampling to H_d true tests so that all profiles (of any discrimination capacity) are within the realms of being practically demonstrated to adhere to Turing's lemma [6]. This is an important ability to possess for model validation, particularly with regards to highly sophisticated DNA evidence interpretation systems.

2. Theory

Importance sampling biases the simulation process so that some elements are chosen more often than at random, and then readjusts for the bias after the simulation. The topic of importance sampling often arises in situations where we want to estimate the probability of a rare event. Importance sampling solves this problem by sampling from an importance density and reweighting the sampled observations accordingly. In general, if X is a random variable with probability density function$p(x)$, and $f(X)$ is some function of X, then the expected value of $f(X)$ is

$$\mathrm{E}[f(X)] = \int_{-\infty}^{+\infty} f(x)p(x)dx$$

If $h(x)$ is also a probability density function which is greater than or equal to zero for the same range of values as $p(x)$ (that is it lies within the support of $p(x)$), then this integral can be rewritten as

$$\mathrm{E}[f(X)] = \int_{-\infty}^{+\infty} f(x)\frac{p(x)}{h(x)}h(x)dx$$

This statement is not very interesting in itself. After all it is equivalent to multiplication by one. However, it is the "trick" which underlies important sampling. If we take a large sample of size S from the *importance density*$h(x)$, then this integral can be approximated by

$$\mathrm{E}[f(X)] \approx \frac{1}{S}\sum_{i=1}^{S} w_i f(x_i)$$

where $w_i = p(x_i)/h(x_i)$ are the *importance weights*. The idea behind importance sampling is that the importance density $h(x)$, can be easier to sample from than the original density, $p(x)$, and yields a low-variance estimate of the desired expectation. The choice of $h(x)$ is somewhat arbitrary, but does dictate the efficiency of the sampling scheme. The process of choosing a good importance distribution is known as *tuning* and can often be very difficult. One might think of this process over-sampling the events of interest, and then *down-weighting* or *biasing* the sample values with weights that reflect the relative probabilities of the events in the importance and original densities. We provide a simple example of importance sampling in Appendix A.

2.1. Application of importance sampling to $\mathrm{H_d}$ true tests

In the problem at hand we might regard X as the *LR*, and $f(X) = X$. That is f is the identity. For each of the 'y' $\mathrm{H_d}$ true tests carried out we calculate a weight, which we call a bias and denote b_y. Here, b_y reflects the size of the bias that leads to the choice in test y. In words, the bias term is the ratio of the probability of the choice using an unbiased method to the probability of that choice had the biasing method been employed. An approximation of the average *LR* (over the Y tests) that would have been obtained had a naïve simulator been used is then:

$$\overline{LR} = \frac{1}{Y}\sum_{y} LR_y b_y \tag{1}$$

and the number of simulations (I) that this would have required had a naïve simulator been used can be approximated by (see Appendix B for derivation):

$$I = \frac{\sum_{y} LR_y}{\overline{LR}} \tag{2}$$

In our single source example from earlier, imagine that we had run one H_d test. The probability of choosing the one genotype that would give a non-zero *LR* using the biased method is one, and this would yield an *LR* of one billion. The probability of choosing this genotype given the unbiased method is 1 in one billion and so $LR_1 = 1 \times 10^9, b_1 = 1 \times 10^{-9}$ and $\overline{LR} = \frac{1}{1}\left(1 \times 10^9\right)\left(1 \times 10^{-9}\right) = 1$. The approximate number of iterations that this corresponds to using a naïve simulator is $I = \frac{1 \times 10^9}{1} = 1 \times 10^9$. This is exactly aligned with our initial expectations, outlined in the introduction. In many instances $\overline{LR} \approx 1$, simplifying Eq. (2) to $I \approx \sum_{y} LR_y$.