Forensic Science International: Genetics 5 (2011) 512–516

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations

James M. Curran [a,*], John S. Buckleton [b]

[a] Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand
[b] ESR, Private Bag 92021, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 4 June 2010
Received in revised form 14 October 2010
Accepted 30 November 2010

Keywords:
DNA
Forensic
Sampling uncertainty

ABSTRACT

There is a variety of methods for assessing sampling uncertainty in likelihood ratio calculations in DNA casework. Sampling uncertainty arises because all DNA statistical methods rely on a database of collected profiles. Such databases can be regarded as a sample from the population of interest. The act of taking a sample incurs sampling uncertainty. In some circumstances it may be desirable to provide some estimate of this uncertainty. We have addressed this topic in two previous publications [1,2]. In this paper we reconsider the performance of the methods using 15 locus Identifiler™ profiles, rather than the 6 locus data used in [1]. We also examine the differences in performance observed when using a uniform prior versus a $1/k$ prior in the Bayesian highest posterior density (HPD) method of Curran et al. [1].

1. Introduction

Sampling uncertainty (also referred to as sampling error by statisticians) arises out of the process of taking a sample from a much larger population. A sample is (usually) much smaller than the population. Therefore, by taking a sample, information will be lost. This phenomenon occurs in the statistical interpretation of DNA evidence. All credible DNA statistics rely, inter alia, on a set of allele frequencies which are estimated from a DNA database which may have been collected explicitly for this purpose, or may be an offender database. A DNA database can be regarded as a sample, which is assumed to be representative, from a much larger population. All DNA statistical calculations, whether they are match probabilities, Random Man Not Excluded (RMNE), paternity indices (PI) or likelihood ratios (*LR*) have uncertainty associated with them from a number of sources including, but not restricted to, sampling uncertainty. This fact has been recognized in the forensic community for some time, and in a number of jurisdictions has been incorporated into active casework [3].

Brenner has argued that an assessment of sampling uncertainty does not assist the court [4]. Whilst his argument has some force it is expected that all forms of uncertainty in evidence will be disclosed to a court. This applies to such disparate evidence types as eyewitness evidence and, we believe, DNA statistics. Incorporation of an assessment of sampling uncertainty is almost ubiquitous in scientific work and it is difficult to see how forensic science should be exempt. However, incorporation of sampling uncertainty into routine testimony is not universal with notable exceptions including the FSS and LGC in the UK.

If a sampling uncertainty correction is to be applied it is desirable that it operates "as advertised." This means that, say, a 95% confidence interval should include the true value 95% of the time. This is referred to as the size of the method.

Our particular interest in this paper is two-fold. In the first instance we wish to revisit previous work assessing the size of various methods [1] but using the full Identifiler™ profiles now available rather than the six locus SGM set investigated previously. Secondly, we are interested in the effect of the prior used in the Bayesian HPD method. Triggs and Curran investigated this to some extent in [2] and recommended the $1/k$ prior. However, the uniform prior, has a powerful natural interpretation whereas the $1/k$ does not. We take the opportunity to investigate this recommendation further here.

2. Methods

The methods for assessing sampling uncertainty under examination are as follows:

1. The "factor of 10" rule
2. The normal approximation
3. The size bias correction

* Corresponding author.
E-mail addresses: j.curran@auckland.ac.nz, curran@stat.auckland.ac.nz (J.M. Curran), john.buckleton@esr.cri.nz (J.S. Buckleton).

1872-4973/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2010.11.007

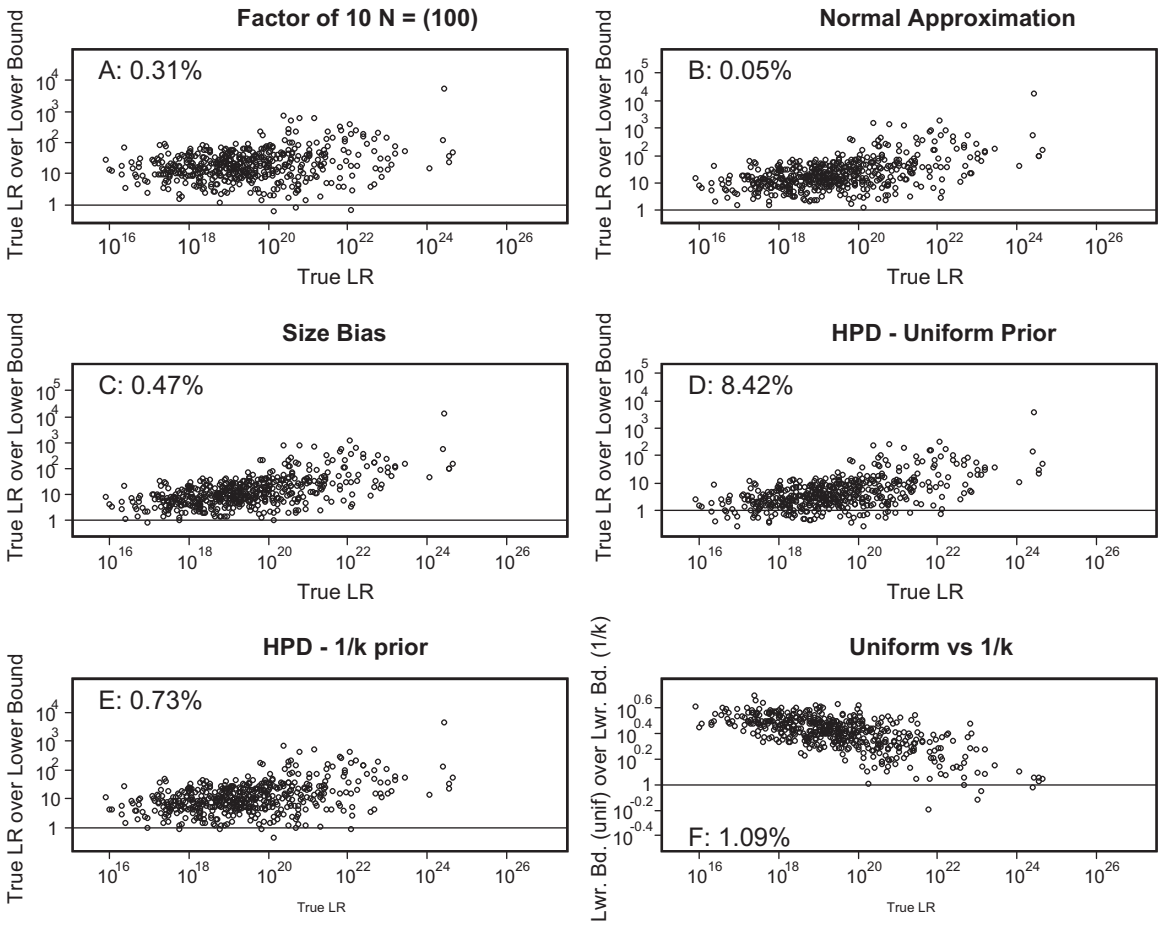

Fig. 1. Simulation results for $N = 100$.

4. The Bayesian HPD with a uniform prior
5. The Bayesian HPD with a $1/k$ prior

We explain each of these methods briefly.

2.1. The "factor of 10" rule

The "factor of 10" rule appears as an un-numbered suggestion in the second National Research Council Report on the forensic evaluation of DNA evidence [5]. This report is usually referred to as NRC II in the literature. The relevant paragraph (p. 160) from the NRC II is as follows:

> "The empirical studies show that the differences between the frequencies of the individual profiles estimated by the product rule from adequate subpopulation databases, (at least several hundred persons) are within a factor of about 10 of each other and that provides a guide to the uncertainty in the determination for a single profile."

As previously noted this method is easily implemented and has some empirical support, but it suffers from the major short coming that it neither reflects the reduced uncertainty that may be obtained by increasing the size of the database nor the increased uncertainty that results from more loci.

2.2. The normal approximation

Chakraborty et al. [6], using the theory of Good [7] suggested a method which relies on a normal approximation, in that the sum of the logarithm of the genotype frequencies has an approximately normal distribution because of the central limit theorem. This method appeared as equations 5.8b and 5.8c in the NRC II. The reader is referred to Buckleton et al. [8] for details of this method.

2.3. The size bias correction

The size bias correction follows the reasoning of Balding [9], later corrected in Evett and Weir [10]. In fairness to Balding, his method was never meant to provide a way to assessing sampling error. The Balding formulae give the Bayesian posterior mean of the allele probabilities. However, because of misinterpretations, the size bias method has substantial uptake in the forensic community as a method for simultaneously dealing with rare or previously unobserved alleles and sampling uncertainty. The method, as implemented, uses the formulae

$$P_{AB} = 2\frac{(x_A + 2)(x_B + 2)}{(2N + 4)(2N + 4)} \quad (1)$$

for heterozygotes and

$$P_{AA} = \left(\frac{x_A + 4}{2N + 4}\right)^2 \quad (2)$$

for homozygotes where N is the number of people in the database, and x_A and x_B are the observed counts of alleles A and B respectively in the database. These formulae have the intuitive explanation that "the suspect plus the individual who left the crime scene stain have been added to the database."