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

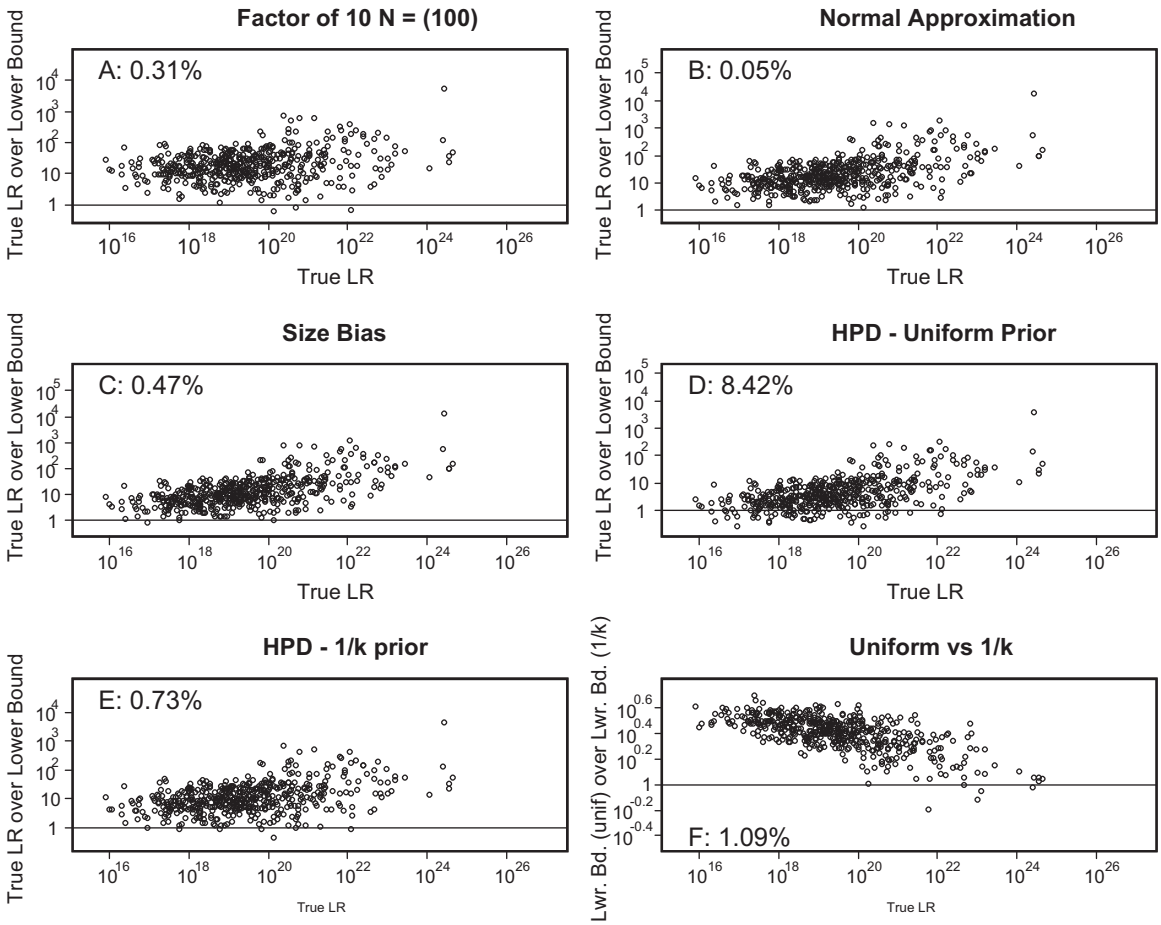

Fig. 1. Simulation results for $N = 100$.

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."

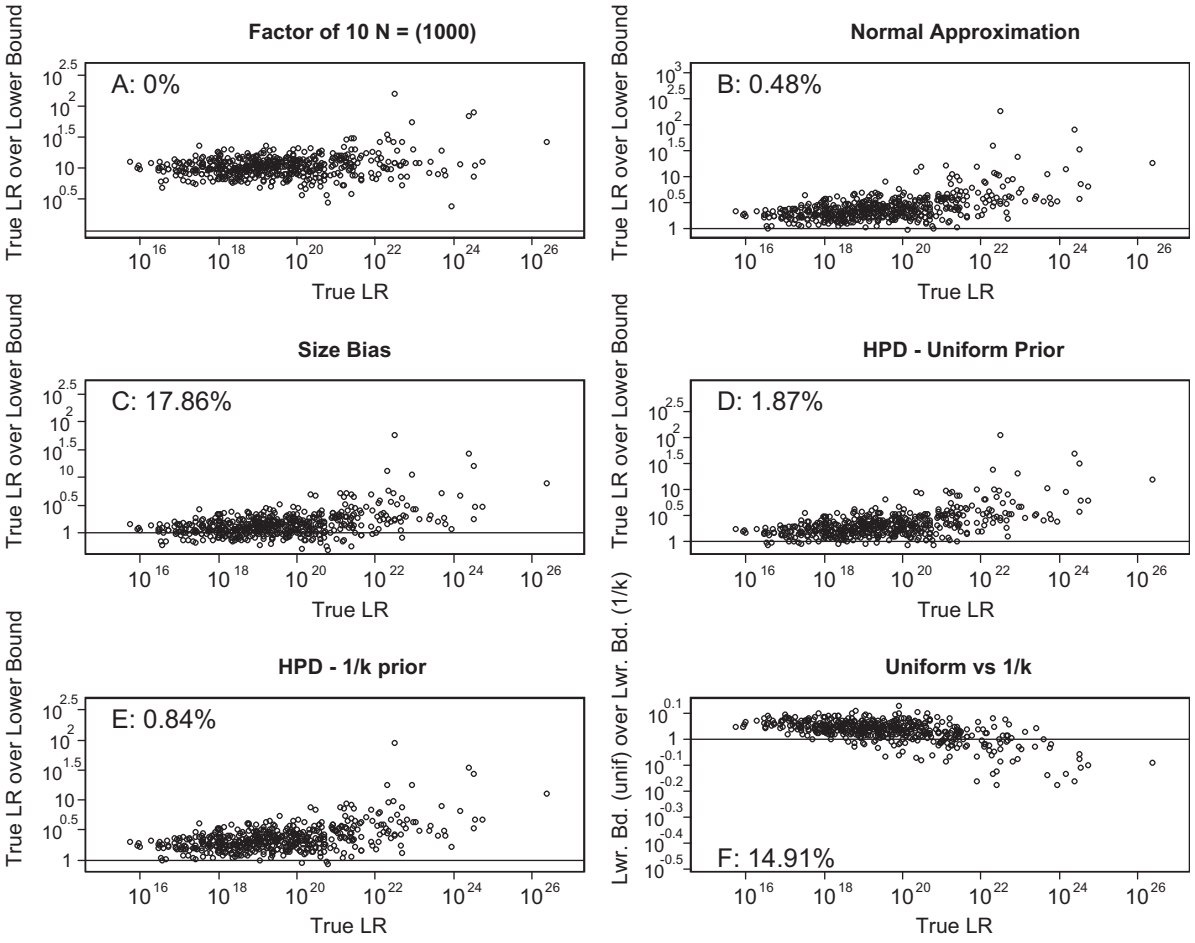

Fig. 2. Simulation results for $N = 1000$.

2.4. The Bayesian HPD

The Bayesian highest posterior density method (HPD) was proposed by Curran et al. [1] following a suggestion in Balding [9]. The idea follows standard Bayesian estimation theory. For each locus a prior probability distribution for every allele possible at that locus is proposed. This prior is updated with the allele count data from a database to yield the posterior density of the allele probabilities given the data. The sampling uncertainty in any DNA statistic may be assessed by taking repeated samples from these posterior allele probability distributions and calculating the statistic. The distribution of this statistic is the posterior sampling distribution of the statistic, and its empirical quantiles can be used to provide a "region of highest posterior density" or credible interval. This allows the user to state that "I'm 99% sure that the statistic is greater than y" where y is a lower bound (0.01 quantile) from the sampling distribution. The method was initially proposed with a uniform prior. That is, in the absence of any data collected, we regard every allele as being equally likely to occur. Whilst this might seem a sensible thing to do, it is not strictly an "uninformative prior" in that it may have an undesirable influence on the posterior distribution of the statistic. One way in which this can be seen is if one considers the posterior mean of an allele A which has been observed x times in a database of size N people (and $2N$ alleles). If the locus at which x occurs has k possible alleles then the expected posterior probability of allele A given the data is

$$E[\pi_A|x] = \frac{x+1}{2N+k}$$

The upshot of this is that although the observed count x is increased by 1, the database size is increased by an amount that reflects the number of alleles at a locus.

Triggs and Curran [2] countered this issue with the introduction of a prior from the same Dirichlet family, but with parameters $1/k$ where k is the number of possible alleles. The expected probability of allele A given the data using this prior is

$$E[\pi_A|x] = \frac{x+(1/k)}{2N+1}$$

The effect of such a prior is smaller in that it moves the posterior expected value less from the maximum likelihood estimate from the data. However the intuition associated with it is less attractive. The $1/k$ prior puts higher probability on the extremes of the distribution rather than the centre. This is equivalent to saying that "allele frequencies will tend to be either fairly small or very large,

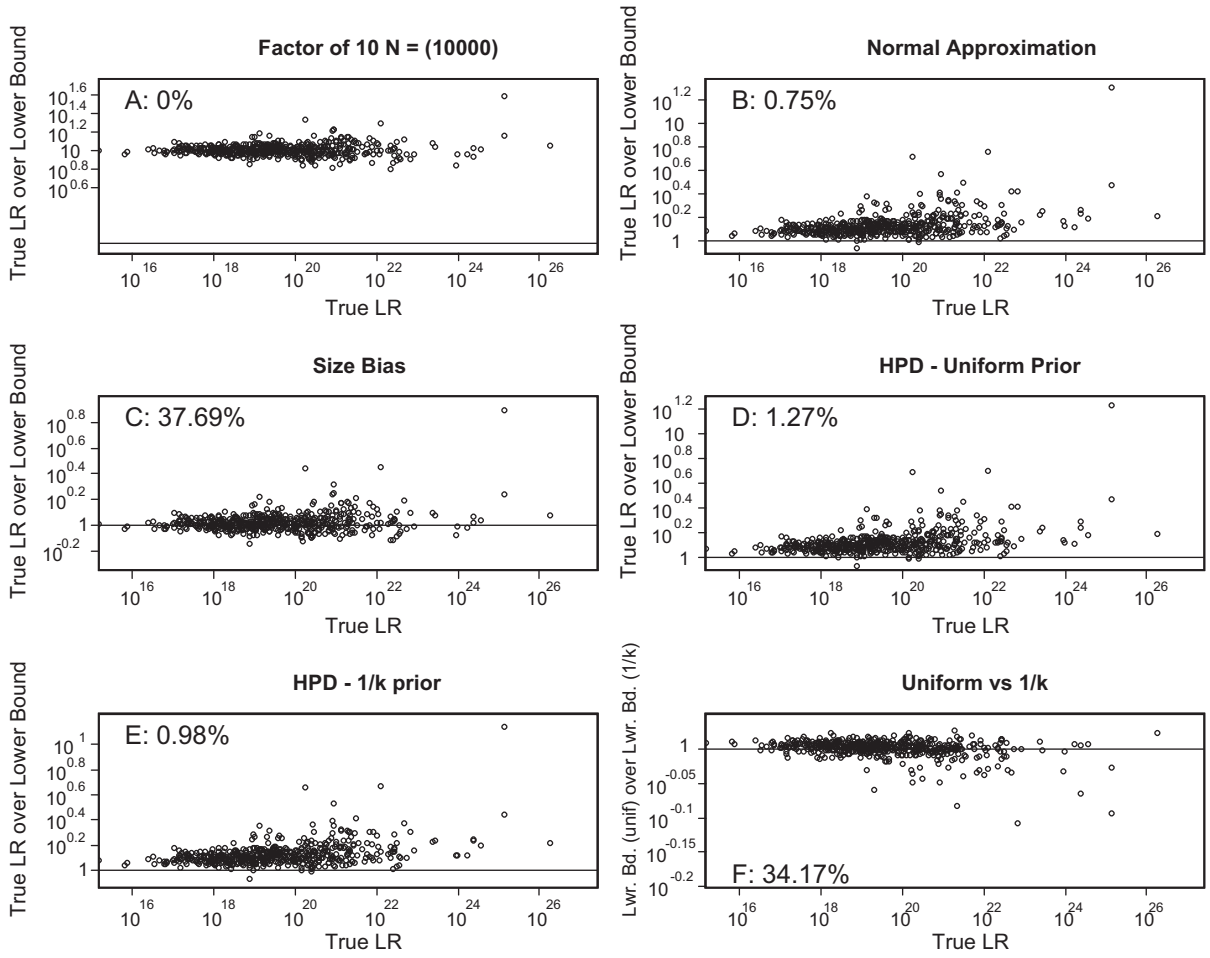

Fig. 3. Simulation results for $N = 10,000$.

but are unlikely to medium values." When regarded from this perspective, we can at least justify it in terms of the "small values" because for most modern multiplexes the allele frequencies are less than about 0.3.

3. Simulation setup

We wished to replicate the work of Curran et al. [1,2] using an Identifiler™ database. To that end we used profile data on 15,545 self declared Caucasians from the NZ DNA database (NZDNADB). In this kind of simulation, it is desirable to have some sort of "truth" against which to compare the results. Therefore, we treated the frequencies from the NZDNADB as being the population allele probabilities. In each iteration:

1. We used these population probabilities to simulate 100 databases of size N with no substructure. Since we have simulated no substructure there is no need for a correction for substructure using, for example, the coancestry coefficient θ. Hence the product rule which assumes of Hardy–Weinberg and Linkage equilibrium is appropriate for this simulation to calculate the profile probability (f). The $LR = 1/f$.
2. The true LR was calculated using the product rule and the true population allele probabilities (assumed from the 15,545 self declared Caucasians from the NZ DNA database (NZDNADB).
3. To test each method we used the product rule and then applied the relevant method using the simulated database allele frequencies.
4. For each database we simulated 100 random profiles again using the database allele frequencies and calculated
   a. The factor of 10:
      i. The factor of 10 LR was calculated as 1/10 times the LR estimated using the product rule and the simulated database frequencies.
   b. The normal approximation:
      i. This was calculated as the 99% one sided lower confidence interval of the LR using the database frequencies, the product rule, and the formulae Chakraborty et al. [6]
   c. The size bias correction:
      i. This was calculated using the product rule and the database frequencies modified using Eqs. (1) or (2).
   d. The Bayesian HPD with a uniform prior
      i. This was calculated as the 99% one sided lower probability interval of the LR using the product rule and the 1000 samples from the posterior allele probability distributions derived from the database frequencies and a uniform prior.
   e. The Bayesian HPD with a $1/k$ prior
      i. This was calculated as the 99% one sided lower probability interval of the LR using the product rule and the 1000 samples from the posterior allele probability distributions derived from the database frequencies and a $1/k$ prior.

This simulation yielded 10,000 values of each statistic for each of the three database sizes we used $N \in \{100, 1000, 10000\}$.

4. Results

Figs. 1–3 display the results from the simulation. Note that each plot only displays a sample of 500 points. This does not affect the interpretation.

The percentage given in the top left of each panel is 100 minus the estimated size and is calculated from the full 10,000 statistics not the 500 selected for use in the plots. The simulations produced 1-sided intervals which should contain the true value 99% of the time, and fail to contain the true value 1% of the time. Therefore, a method with a size substantially larger than 1% indicates that the method does not have the desired properties.

Each first five of the panels A–E shows the ratio of the true value of the *LR* divided by the estimate of the *LR*. In the case of the normal approximation and the two highest posterior density methods this is the lower bound. In the case of the factor of 10 and the size bias method this is not strictly a lower bound, although it is not exactly clear what term should be used. If the ratio is greater than one, then this means the estimated lower bound is lower than the true value, as desired. The ratios are plotted with respect to the true *LR* so that the effect of the method on rare and common profiles can be observed.

Subjectively the most desirable outcome is a strip of points just above 1. This would mean that the lower bound was conservative but not excessively so. We accept the subjective nature of this judgment and point out that it applies to criminal casework but not to civil casework.

It can be seen from the plots that the ratio decreases as N increases for all methods with the exception of the factor of 10 (see Figs. 1A, 2A, and 3A). This behavior is expected as the factor of 10 method does not reflect the database size.

In most of the plots a trend towards greater values of the ratio and to greater variability can be perceived as the true *LR* decreases. This is expected and is akin to the trend as N increases. For rare profiles the sampled count of a few alleles may be very low and hence the intervals very wide and less well centered.

We note that as the database size increases the performance of the size bias method worsens with a significant portion of the estimates producing a *LR* that is bigger than the true value (see Figs. 1C, 2C, and 3C). Once again this is expected because the size bias estimator is not designed to deal with sampling uncertainty.

The Bayesian HPD variants perform well, both providing tight lower bounds even with modest ($N \geq 1000$) database sizes (Figs. 1D and E, 2D and E, and 3D and E). It is worth noting that the Bayesian HPD with a uniform prior does not perform well for small databases ($N \leq 100$). Fig. 1D shows that the method has poor coverage probability for $N = 100$. This effect is magnified when the prior allows for a large number of "possible" alleles, even though most of them may not have been observed. The $1/k$ prior does not have this issue. This finding further confirms our previous finding [2] that $1/k$ has more desirable properties. The sixth panel in the bottom right of each plot attempts to address the issue of whether we should prefer a uniform prior or a $1/k$ prior. The plot displays the ratio of the uniform lower bound to the $1/k$ lower bound. If this ratio is greater than one, then this means that the uniform prior has a larger lower bound than the $1/k$ lower bound, i.e. $Ratio > 1 \Rightarrow LB_{1/k} < LB_{Unif}$. Similarly if the ratio is less than one then the uniform prior has a smaller lower bound than the $1/k$ lower bound, $Ratio < 1 \Rightarrow LB_{1/k} > LB_{Unif}$. These are conceptually hard ideas to get ones head around. If situation 1 ($Ratio > 1 \Rightarrow LB_{1/k} < LB_{Unif}$) is the case the majority of the time, then this means that the $1/k$ prior is producing lower bounds that are closer to the true value. If situation 2 ($Ratio < 1 \Rightarrow LB_{1/k} > LB_{Unif}$) is the case the majority of the time, then this means that the uniform prior is producing lower bounds that are closer to the true value. Lower bounds that are close, but not greater than the true value are desirable. For each of the three databases, situation 1 holds about two thirds of the time on average. This means, about two thirds of the time the $1/k$ prior will produce a lower bound that is a) less than the true value as desired, but b) closer to the true value than the uniform prior. This suggests that the $1/k$ prior is preferable.

5. Conclusions

In this paper we repeated the work of Curran et al. [1] and some of the work of Triggs and Curran [2] using full Identifiler™ profiles.

Because of the immense power of DNA minor differences in the *LR* may be immaterial in the view of a court. This paper represents either a confirmation that these differences are small and hence tolerable or a recommendation for best practice depending on the view of the reader.

We found that the factor of 10 rule and the size bias method are unsuitable for assessing sampling error. The normal approximation does well, especially with large databases, but it is routinely out performed by the two Bayesian HPD method. Furthermore, we concur with the findings of Triggs and Curran [2], in that use of a $1/k$ prior with the Bayesian HPD method seems to produce more desirable estimates of sampling uncertainty.

The relative difference between the estimates and the true values was small in all cases and may be tolerable. This difference is reduced for larger databases. The current best practice method for assessing sampling uncertainty of those examined appears to be the Bayesian HPD method with a $1/k$ prior.

Acknowledgements

We gratefully acknowledge the comments of Jo-Anne Bright, Susan Vintiner and two anonymous referees which have greatly improved this manuscript.

References

[1] J.M. Curran, J.S. Buckleton, C.M. Triggs, B.S. Weir, Assessing uncertainty in DNA evidence caused by sampling effects, Sci. Justice 42 (1) (2002) 29–37.
[2] C.M. Triggs, J.M. Curran, The sensitivity of the Bayesian HPD method to the choice of prior, Sci. Justice 46 (3) (2006) 169–178.
[3] P.D. Gill, L.A. Foreman, J.S. Buckleton, C M Triggs, H. Allen, A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations, Forensic Sci. Int. 131 (2003) 184–196.
[4] C H. Brenner, DNA frequency uncertainty – why bother, Available from: www.dna-view.com/noconfid.htm, 1997.
[5] NRC_II, National Research Council Committee on DNA Forensic Science, The evaluation of forensic DNA evidence, Washington, D.C.: National Academy Press, 1996.
[6] R. Chakraborty, M.R. Srinivasan, S F Daiger, Evaluation of standard errors and confidence intervals of estimated multilocus genotype probabilities and their implications in DNA, Am. J. Hum. Genet. 52 (1993) 60–70.
[7] I.J. Good, The population frequencies of species and the estimation of population parameters, Biometrika 40 (3-4) (1953) 237–264.
[8] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, Florida, 2005.
[9] D.J. Balding, Estimating products in forensic identification using DNA profiles, J. Am. Stat. Assoc. 90 (431) (1995) 839–844.
[10] I.W. Evett, B.S. Weir, Interpreting DNA Evidence–Statistical Genetics for Forensic Scientists, Sinauer Associates, Inc., Sunderland, 1998.