On the exact distribution of the number of alleles in DNA mixtures

Torben Tvedebrink - tvede@math.aau.dk
Department of Mathematical Sciences, Aalborg University, Denmark
Section of Forensic Genetics, Department of Forensic Medicine,
Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

Introduction

When more than one individual contributes biological material to a forensic stain, the resulting DNA type is termed a DNA mixture. DNA mixtures occur frequently in forensic genetic casework, and in recent years much research has been devoted to this subject.

This poster presents a derivation of the exact distribution of the number of alleles for any number of profiles and investigated loci. The per locus number of observed alleles is of interest as it indicates the plausible range of the number of contributors. Furthermore, the total number of alleles across all loci are used by some forensic geneticists to estimate the probability that an allele has not been detected.

Recursion – how to obtain α_m

The total number of alleles observed for m contributors, $N(m)$, depends on the number of loci, L, through the locus specific allele counts, N_l, by $N(m) = \sum_{l=1}^{L} N_l(m)$. Hence, we focus on computing the distribution of $N_l(m)$ below.

The expression for $P(N_l(m) = n)$ gets complicated for increased m. Hence, let $\boldsymbol{\alpha}_m$ denote a vector of the numbers of unique alleles that the m DNA profiles carry at a given locus. In the simplest case $m = 1$, the DNA profile may either be heterozygous, $\boldsymbol{\alpha}_1 = (1, 1)$, or homozygous, $\boldsymbol{\alpha}_1 = (2)$. In order to compute $P(N(m) = n)$, the only information necessary is $\boldsymbol{\alpha}_m$.

Extending a set of profiles can happen in a limited number of ways. The added profile may be

- heterozygous, sharing none (h_0), one (h_1) or both (h_2) alleles with the previous profiles;
- homozygous, sharing none (H_0) or one (H_1) allele with the previous profiles.

Mathematically, this can be formulated in terms of updating $\boldsymbol{\alpha}_m$ to obtain $\boldsymbol{\alpha}_{m+1}$, where l is the dimension of $\boldsymbol{\alpha}_m$:

$$\boldsymbol{\alpha}_{m+1} = \begin{cases} \boldsymbol{\alpha}_m + \boldsymbol{e}_i + \boldsymbol{e}_j, & 1 \le i < j \le l & (h_2) \\ \boldsymbol{\alpha}_m + \boldsymbol{e}_i + \boldsymbol{e}_j, & 1 \le i \le l,\ j = l{+}1 & (h_1) \\ \boldsymbol{\alpha}_m + \boldsymbol{e}_i + \boldsymbol{e}_j, & i = l{+}1,\ j = l{+}2 & (h_0) \\ \boldsymbol{\alpha}_m + 2\boldsymbol{e}_i, & 1 \le i \le l & (H_1) \\ \boldsymbol{\alpha}_m + 2\boldsymbol{e}_i, & i = l{+}1 & (H_0) \end{cases}$$

Obtaining $\boldsymbol{\alpha}_2$ from $\boldsymbol{\alpha}_1$ is shown in Fig. 1. For example, in the topmost path (green), the first profile is homozygous, $\boldsymbol{\alpha}_1 = (2)$, and adding a heterozygous profile that shares no allele, h_0, we obtain $\boldsymbol{\alpha}_2 = (2, 1, 1)$.

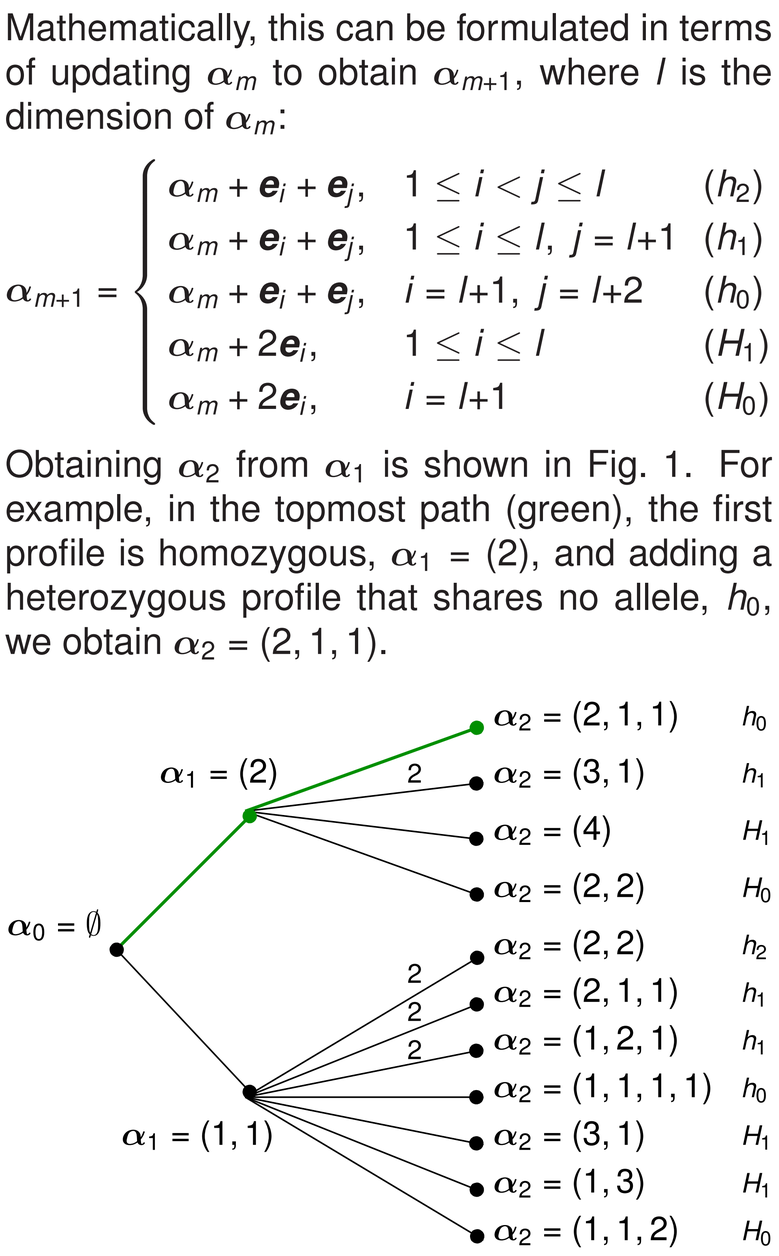

Fig. 1: Recursion showing how to obtain $\boldsymbol{\alpha}_2$ from $\boldsymbol{\alpha}_1$.

Computing $P(N_l(m) = n)$ based on $\boldsymbol{\alpha}_m$

In order to evaluate $P(N_l(m){=}n)$, we must consider the set $\mathcal{A}_m^n$ of all $\boldsymbol{\alpha}_m$-vectors of dimension n. Each of these vectors, $\boldsymbol{\alpha}_m{=}(\alpha_1, \ldots, \alpha_n)$, gives the relevant powers of allele frequencies, which represent the count of the specific allele present in the DNA mixture. For example, for $\boldsymbol{\alpha}_2{=}(2, 2)$ both alleles are represented twice, e.g. both mixtures (A_iA_i,A_jA_j) and (A_iA_j,A_iA_j) are represented by $(2, 2)$.

The fact that more mixtures are represented by the same $\boldsymbol{\alpha}_m$-vector motivates the need for a weight-function, $c(\boldsymbol{\alpha}_m)$. The $c(\boldsymbol{\alpha}_m)$-function counts the number of times $\boldsymbol{\alpha}_m$ is formed by the recursions (see the *Recursion*-box), e.g. $\boldsymbol{\alpha}_2 = (2, 1, 1)$ is formed twice in Fig. 1. Furthermore, if the added profile is heterozygous of type h_1 or h_2 (sharing one or both alleles), the formed $\boldsymbol{\alpha}_m$-vector is counted twice (due to unordered observations). Hence, to obtain $P(N_l(m) = n)$, we evaluate

$$P(N_l(m) = n) = \sum_{\boldsymbol{\alpha}_m \in \mathcal{A}_m^n} c(\boldsymbol{\alpha}_m) \sum_{i_1,\ldots,i_n}^{\neq} p_{i_1}^{\alpha_1} p_{i_2}^{\alpha_2} \ldots p_{i_n}^{\alpha_n},$$

where $i_1, \ldots, i_n$ are n different indices, and $p_{i_1}, \ldots, p_{i_n}$ are the allele frequencies of locus l.

Plots of some results

Fig. 2 shows the distributions of $P(N(m) = n)$ for $m = 1, \ldots, 8$ of the ten SGM Plus loci, $L = 10$.

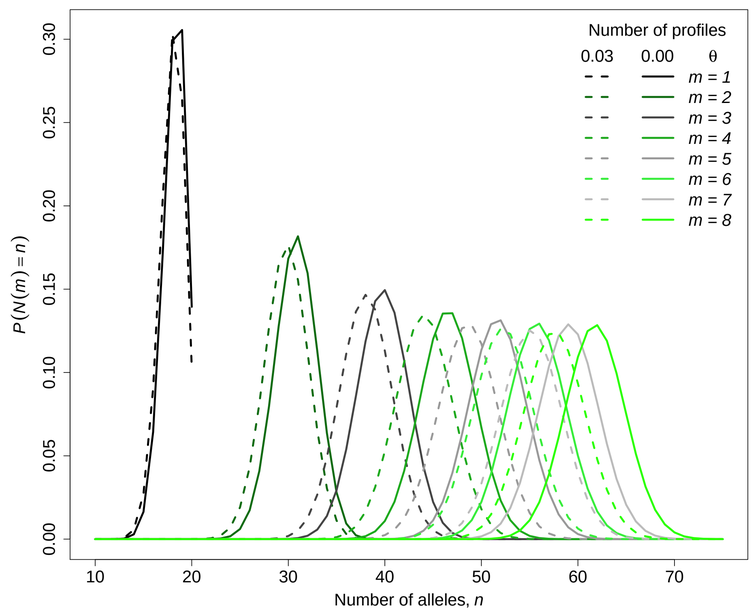

Fig. 2: $P(N(m) = n)$ of ten SGM Plus loci.

Using Bayes' theorem, we obtain the posterior for the number of contributors, $P(m|n) = P(N_l(m) = n)P(m)/P(n)$.

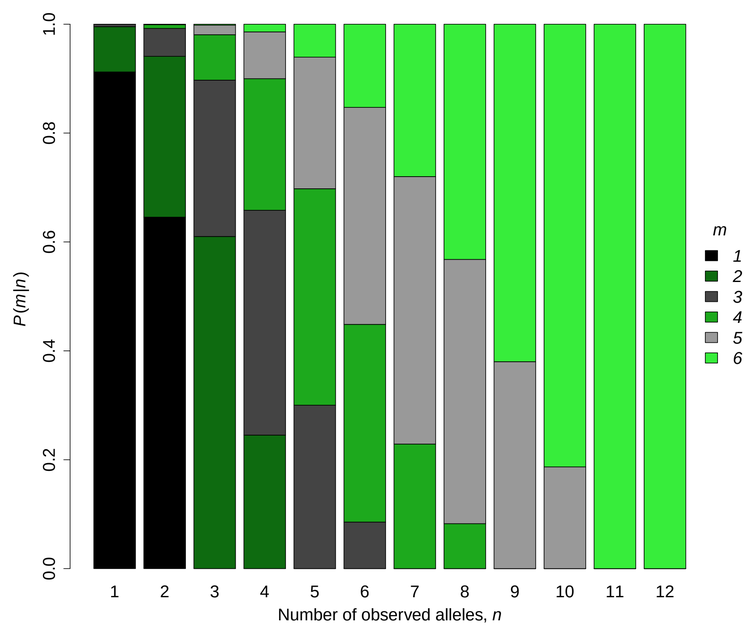

Fig. 3: Distribution of the number of contributors, m, given that n alleles are observed, $P(m|n)$, for vWA.

Conclusion

The exact distribution of the number of alleles was derived using recursion relations.

Based on $P(N_l(m) = n)$ and a prior distribution, $P(m)$, the posterior of the number of contributors, $P(m|n)$, can be computed.

It was demonstrated that the drop-out probability based on the number of alleles, $\hat{P}(D)$, was correlated with θ.

Drop-out

Allelic drop-out will cause fewer alleles to be detected in a sample. This led Gill et al. [1] and more recently Haned et al. [2, 3] to use only the number of observed alleles to estimate the drop-out probability.

However, a relatively low number of observed alleles may also be caused by shared ancestry and subpopulation stratification (often modelled by the θ-correction). In Fig. 4, the estimated drop-out probabilities for a two-person DNA mixture based on Monte Carlo simulations [1] are plotted against θ.

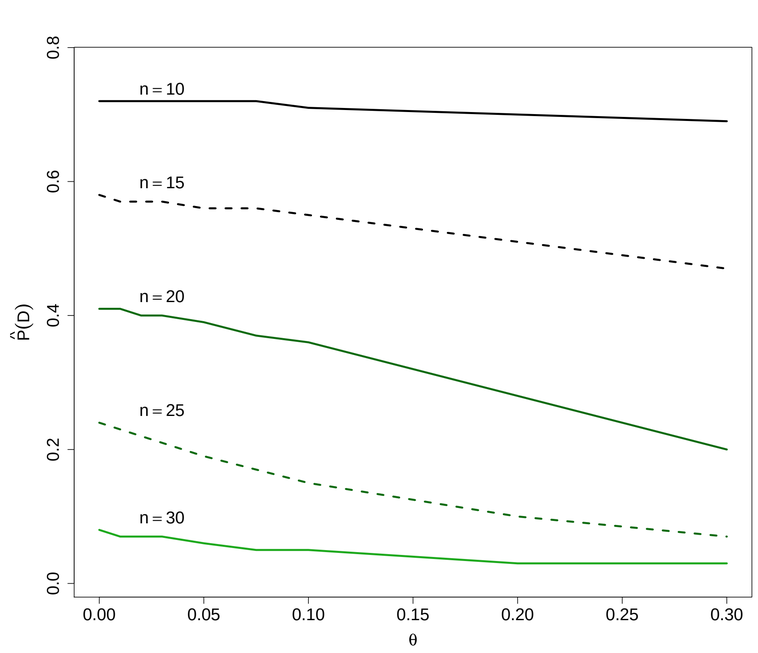

Fig. 4: Monte Carlo based estimate of $P(D)$ plotted against θ for various numbers of observed alleles, n.

Fig. 4 shows that the estimated drop-out probability, $\hat{P}(D)$, is correlated with θ. The declining trend shows that the Monte Carlo approach will overestimate $P(D)$ when $\theta > 0$. For e.g. $n = 25$, the estimates are $\hat{P}(D; \theta = 0.05) = 0.19$ and $\hat{P}(D; \theta = 0) = 0.25$, respectively.

References

[1] Gill, P., A. Kirkham, and J. Curran (2007). *LoComatioN: A software tool for the analysis of low copy number DNA profiles*. Forensic Sci Int 166(2-3): 128-138.

[2] Haned, H. (2011). *Forensim: An open-source initiative for the evaluation of statistical methods in forensic genetics.* Forensic Sci Int Genet 5(4): 265-268.

[3] Haned, H., K. Slooten, and P. Gill (2012). *Exploratory data analysis for the interpretation of low template DNA mixtures*. Forensic Sci Int Genet 6(6): 762-774.