GHEP forensic exercise 2023, advanced level

Thore Egeland and Magnus Dehli Vigeland

General instructions

This is a multiple-choice test consisting of 20 questions. For each question exactly one alternative is correct. There may be issues related to for instance rounding, so if your answer does not agree precisely with any alternative, you should choose the closest option.

Throughout we make the following assumptions unless stated otherwise:

- no deviations from Hardy Weinberg Equilibrium,
- all profiles are from unrelated individuals typed for independent, autosomal markers,
- no drop-out, drop-in, silent alleles or mutations.

For *a given locus*, the genotypes of two individuals are said to be a

- *full match* if the genotypes are identical (e.g., a/b - a/b, or a/a - a/a),
- *partial match* if one allele is identical and the other differs (e.g., a/b - a/c, or a/b - a/a),
- *mismatch* if both alleles differ (e.g., a/b - c/c).

For any given pair of full profiles containing L loci there will be x loci with full matches, y with partial matches, and $L-x-y$ mismatches.

We use the notation 1.2e-4 to mean $1.2*10^{-4} = 0.00012$.

Most problems can be solved using paper, pencil, and a calculator. However, you are free to use software whenever possible. Some comments on software and references appear at the end.

The best of luck!

Questions

Match probabilities and database search

For questions 1 - 3 we assume that two unrelated individuals are randomly chosen from the population and typed for one SNP marker, with allele frequencies 0.8 and 0.2.

1) The probability of a full match is
    a) 0.1024
    b) 0.2176
    c) 0.3200
    d) 0.4112
    e) 0.5136

2) The probability of a partial match is
    a) 0.2176
    b) 0.2217
    c) 0.3264
    d) 0.4352
    e) 0.5376

3) The probability of a mismatch is
    a) 0.0256
    b) 0.0512
    c) 0.1024
    d) 0.3400
    e) 0.6800

Now we consider STR markers. For simplicity we assume that the allele frequencies are the same for all loci. Suppose that for each locus, the probabilities of mismatch, partial match and full match are, respectively, p_0, p_1, and p_2. Assume two unrelated individuals are chosen randomly from the population and typed at L loci.

4) Assume $p_0 = 0.66$, $p_1 = 0.32$, and $p_2 = 0.02$. If $L = 10$, the probability that all markers show either a partial or a full match, is
    a) 1.024e-17
    b) 1.126e-05
    c) 2.064e-05
    d) 0.0157
    e) 0.3400

5) Assume that all markers show either partial or full match. We can with certainty exclude
    a) that they are unrelated
    b) that they are full siblings
    c) that they are related as parent-child
    d) that they are dizygotic twins
    e) none of the above

6) The probability of a full match at x loci and a partial match at y loci, where $0 \le x + y \le L$, is
    a) $p_0^{L-x-y}\ p_1^{y} p_2^{x}$
    b) $\frac{L!}{(L-x-y)!y!x!}\ p_1^{y} p_2^{x}$
    c) $\frac{(x+y)!}{y!x!} p_1^{y}\ p_2^{x}$
    d) $\frac{L!}{(L-x-y)!y!x!} p_0^{L-x-y}\ p_1^{y} p_2^{x}$
    e) $\frac{L!}{(L-x-y)!y!x!}(p_0^{L-x-y} +\ p_1^{y} + p_2^{x})$

Consider a database of 10 000 DNA profiles from unrelated individuals typed at the markers described previously. All pairs of profiles are compared.

7) The number of pairwise comparisons is
   a) 10 000
   b) 20 000
   c) 49 995 000
   d) 99 990 000
   e) 100 000 000

8) The expected number of pairs with full or partial match at all 10 loci is
   a) 0
   b) 1
   c) 1032
   d) 2064
   e) 4128

We next consider a realistic dataset comprising $L = 15$ forensic STR markers. We have simulated $N = 10\,000$ profiles from unrelated individuals, all males, using the frequencies of the 15 loci in Identifiler_Spain.csv. The results are summarized in the Table 1 below.

		Partial matches															
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Full matches	0	110	1927	17065	91155	331068	877112	1746402	2658370	3120650	2816266	1941577	1005938	377020	97399	15453	1081
	1	317	5513	44158	214900	706337	1671787	2937890	3904660	3934796	2994392	1693453	686447	190737	32274	2496	0
	2	423	7231	51142	223967	656930	1378643	2118666	2417941	2053746	1278114	566904	170209	30831	2537	0	0
	3	399	5502	35097	137126	355033	648927	859636	824837	572215	281287	91553	18079	1593	0	0	0
	4	208	2768	15874	54209	122920	193941	216289	170186	93179	33231	7210	708	0	0	0	0
	5	97	988	4878	14896	28828	38184	35002	21472	8504	2040	211	0	0	0	0	0
	6	24	267	1140	2978	4689	4932	3655	1666	446	52	0	0	0	0	0	0
	7	5	41	185	366	528	467	202	64	6	0	0	0	0	0	0	0
	8	1	9	19	38	43	18	7	2	0	0	0	0	0	0	0	0
	9	0	3	1	0	4	1	0	0	0	0	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1 Number of comparisons that show partial and full matches. Rows and columns show the number of markers with full or partial match, respectively. For instance, 267 comparisons give full matches in 6 markers, a partial match in 1 marker, and hence mismatches in 8 markers.

For the two next questions we consider a comparison between two randomly chosen unrelated individuals. Base your answers on Table 1.

9) The most likely outcome is
   a) 1 full match and 8 partial matches
   b) 8 full matches and 1 partial match
   c) 8 full matches and 2 partial matches
   d) 5 mismatches and 5 partial matches
   e) 8 mismatches and 5 partial matches

10) The probability of a full match or a partial match at all 15 loci, is
    a) 2.162e-05
    b) 1.737e-04
    c) 0.0787
    d) 0.1081
    e) 0.8686

A male *case* profile, left by the POI (Person Of Interest), is obtained from a crime scene. This profile is searched against the database. The result is:

Full matches \ Partial matches	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	10	44	108	228	423	505	490	362	202	90	17	0	0
1	0	0	2	33	116	272	519	756	799	616	376	166	44	7	0	0
2	0	1	6	32	129	303	441	551	519	347	136	30	9	2	0	0
3	0	0	10	29	79	151	238	223	148	79	27	4	1	0	0	0
4	0	0	3	15	40	58	51	53	24	8	3	0	0	0	0	0
5	0	0	1	8	10	16	7	7	6	1	0	0	0	0	0	0
6	0	0	1	1	2	1	1	1	1	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2 The result of searching a case profile against the database of 10 000 profiles. For instance, there are 272 profiles in the database that show 1 full match and 5 partial matches.

11) Based on this, it is reasonable to conclude that POI
    a) is not in the database
    b) cannot be the first cousin of anyone in the database
    c) cannot be the brother of anyone in the database
    d) cannot be the son of anyone in the database
    e) is in the database

Next, denote by A_1, …, A_N the individuals in the database, where N = 10 000. We will follow the approach of Slooten and Meester (2014) to calculate the posterior probability that each A_i is related to POI.

Let π_0 denote the prior probability that POI has no relative in the database. Furthermore, we assume that POI has at most one relative in the database, and that all A_i's have the same prior probability of being this relative. Let r_i denote the LR comparing "there is a parent-child relationship between A_i and POI" to "POI and A_i are unrelated". According to Proposition 1 (*op. cit.*), the posterior probability that A_i is the relative of POI is,

$$\frac{(1-\pi_0)r_i}{\pi_0 N + (1-\pi_0)\sum_{j=1}^{N} r_j}.$$

From Table 2 we can identify three individuals that could potentially have a parent-child relationship to POI. Let these be A_1, A_2 and A_3. You are told that the corresponding LRs are r_1 = 1533, r_2 = 716 and r_3 = 148. All other r-s are 0 as a consequence of the assumptions made initially.

12) Assume $\pi_0 = 0.5$. The probabilities that A_1, A_2 and A_3 are in a parent-child relationship to POI are, respectively,
    a) 0, 0, 0
    b) 0.124, 0.058, 0.012
    c) 0.153, 0.072, 0.015
    d) 0.640, 0.299, 0.062
    e) 1, 0.467, 0.097

13) Let $S = \sum_{j=1}^{N} r_j = 2397$. The probability that *one* of the individuals A_1, A_2, A_3 is in a parent-child relationship to POI is
    a) $\frac{(1-\pi_0)}{\pi_0 N+(1-\pi_0)S}$
    b) $\frac{S}{\pi_0 N+(1-\pi_0)S}$
    c) $\frac{(1-\pi_0)S}{\pi_0 N+(1-\pi_0)}$
    d) $\frac{(1-\pi_0)S}{\pi_0 N+(1-\pi_0)S}$
    e) $\frac{S}{S+1}$

14) The POI was not found after having investigated the families of individuals A_1, A_2 and A_3. The investigator decided to check the families of the individuals in the database that shared at least one allele for precisely 14 markers. The number of such individuals is
    a) 16
    b) 17
    c) 20
    d) 25
    e) 26

Number of mixture contributors

The following questions address one way of estimating the number of contributors to a mixture. If a DNA mixture has *c* contributors, it is possible to observe anything between 1 and 2*c* alleles at a given marker. Of particular interest is the probability that a *c* +1 person DNA mixture is *misclassified* as a mixture coming from no more than *c* individuals. That is: *what is the probability that c+1 persons have at most 2c different alleles among them for all markers*? Below we estimate the number of contributors by the minimum number of individuals needed to explain the mixture. Table 3 shows the probabilities of seeing 1-6 different alleles for each marker when there are three contributors. For instance, the probability that there will be 4 different alleles for D8S1179 is 0.4439.

	1	2	3	4	5	6
D8S1179	0.0008	0.0426	0.2694	0.4439	0.2170	0.0263
D21S11	0.0005	0.0282	0.2075	0.4284	0.2839	0.0515
D7S820	0.0006	0.0394	0.2872	0.4665	0.1903	0.0160
CSF1PO	0.0020	0.1278	0.5217	0.3098	0.0378	0.0009
D3S1358	0.0006	0.0506	0.3660	0.4721	0.1090	0.0018
TH01	0.0007	0.0567	0.3810	0.4603	0.0998	0.0015
D13S317	0.0013	0.0665	0.3387	0.4293	0.1523	0.0117
D16S539	0.0017	0.0919	0.4220	0.3906	0.0896	0.0043
D2S1338	0.0002	0.0150	0.1456	0.4020	0.3539	0.0833
D19S433	0.0018	0.0709	0.3309	0.4138	0.1646	0.0180
VWA	0.0006	0.0397	0.2878	0.4662	0.1902	0.0156
TPOX	0.0135	0.2377	0.4872	0.2367	0.0246	0.0002
D18S51	0.0001	0.0104	0.1350	0.4099	0.3621	0.0825
D5S818	0.0042	0.1745	0.5164	0.2730	0.0312	0.0007
FGA	0.0001	0.0116	0.1481	0.4299	0.3442	0.0661

Table 3 The probabilities that a 3-person mixture shows 1,2, ..., 6 different alleles for each marker.

15) Consider the marker D8S1179. The probability that a 3-person mixture will misclassified is
    a) 0
    b) 4.08e-06
    c) 0.313
    d) 0.500
    e) 0.757

16) Consider all markers. The probability that a 3-person mixture will be misclassified is
    a) 8.84e-07
    b) 0.026
    c) 0.500
    d) 0.797
    e) 1.000

17) If theta correction is considered, the probability of misclassification will
    a) remain unchanged
    b) be larger
    c) be smaller
    d) be 0
    e) be 0.01

The remaining exercises relate to the app http://apps.math.aau.dk/noa/ made by Torben Tvedebrink.

18) For this exercise you should use the app with input file Identifiler_Spain.csv. The probability that a four-person mixture is misclassified is
    a) 0
    b) 0.04
    c) 0.5
    d) 0.7
    e) 1

Figure 1 below was made by running the app with 2, 3, 4, 5 and 6 contributors to the mixture. The output was downloaded from the app and plotted in R. The figure shows the distributions of the total number of different alleles based on all markers, for mixtures with 2-6 contributors.

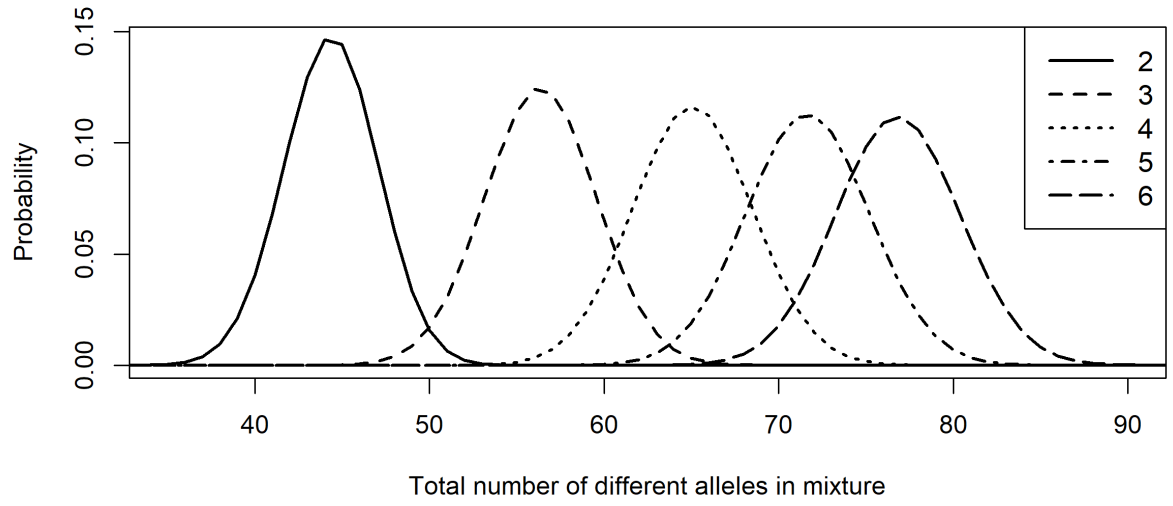

Figure 1 Distribution for the total number of different alleles.

19) Assume the number of different alleles from three independent mixture cases are respectively 45, 55, and 65. The most likely number of contributors to these mixtures are, respectively,
    a) 2, 2, and 4
    b) 2, 3, and 4
    c) 2, 3, and 5
    d) 3, 3, and 4
    e) 3, 3, and 6

20) It is reasonable to conclude that we can distinguish best between mixtures coming from
    a) 2 and 3 contributors
    b) 2 and 4 contributors
    c) 3 and 5 contributors
    d) 4 and 6 contributors
    e) 5 and 6 contributors

References

The solution to these exercises will contain the code used to generate the data. We have used the R libraries `DNAtools` (Tvedebrink et al., 2014), `numberofalleles` (Kruijver and Curran, 2022) and `forrel` (Vigeland, 2021).

1. Kruijver, M, and Curran, JM. "The number of alleles in DNA mixtures with related contributors." *Forensic Science International: Genetics* 61, 2022.
2. Slooten, K, and Meester, R. "Probabilistic strategies for familial DNA searching." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63.3, 2014.
3. Tvedebrink, T, Eriksen, PS, Curran, JM, Mogensen, HS, and Morling, N. "Analysis of Matches and Partial-Matches in a Danish DNA Reference Profile Data Set." *Forensic Science International: Genetics*, 2014.
4. Vigeland, MD. *Pedigree analysis in R*. Academic Press, 2021.