

GHEP mixture exercise 2022, advanced level

Thore Egeland 2022-03-25

General instructions

This is a multiple-choice test containing 20 questions. For each question you are asked to choose the correct alternative. There may be issues related to e.g., rounding and so your answer may differ slightly from the correct one. If your answer does not agree exactly with any alternative, you should choose the closest option.

Throughout we make standard simplifying assumptions including:

- independent autosomal markers
- no mutations
- no silent alleles
- no deviations from Hardy Weinberg Equilibrium (except if not specified otherwise)
- all contributors to mixtures are unrelated

The test consists of three parts, A, B and C. The two first parts can be solved using paper, pencil, and a calculator. However, you can alternatively use software whenever possible. For instance, [LRmix Studio](#) can be used to answer questions 1-6 (but you have to prepare input files) and partly solve other exercises in Part A and B. In parts A and B we use the discrete model to calculate likelihood ratios. Drop-in is not modelled. Drop-out is only modelled if explicitly stated.

Part A

We consider one marker with 10 alleles denoted 1, 2, ..., 10. Each allele has frequency $p = 0.1$. We assume that the person of interest, POI, has genotype 1/2. The alleles (peak heights) in the mixture are

1 (426), 2 (414), 3 (46), 4 (44), 5 (37), 6 (33).

- 1) Assume a peak threshold of 50 is used so that the mixture is {1, 2}. If the discrete model is used, the likelihood ratio comparing the hypothesis H_{p1} : *POI contributed* to H_{d1} : *an unknown contributed* is
 - a. $1/(2p^2)$
 - b. 100
 - c. $1/(2p(1-p))$
 - d. $2p^2$
 - e. 1

2) If threshold of 40 is used the mixture is {1, 2, 3, 4}. Let LR_2 compare the hypothesis H_{p2} : *POI + unknown contributed* to H_{d2} : *two unknowns contributed*. If a threshold of 30 is used the mixture is {1, 2, 3, 4, 5, 6}. Let LR_3 compare the hypothesis H_{p3} : *POI + two unknowns* to H_{d3} : *three unknowns*. If the discrete model is used, we can conclude without calculation that

- a. $LR_1 \leq LR_3 \leq LR_2$
- b. $LR_1 \leq LR_2 \leq LR_3$
- c. $LR_2 \leq LR_1$ and $LR_3 \leq LR_1$
- d. $LR_1 = LR_2 = LR_3$
- e. $LR_1 \leq LR_3$ and $LR_2 = LR_3$

3) The LR_2 in the previous question is

- a. $1/p^2$
- b. $1/(6p^2)$
- c. $1/(24p^2)$
- d. $12p^2$
- e. $1/(12p^2)$

4) The LR_3 in the question 2 is

- a. $1/(30p^4)$
- b. $1/(60p^2)$
- c. $1/(15p^2)$
- d. $1/(30p^2)$
- e. $1/(2p^2)$

5) We use the dropout model described in the supplementary material of Haned, Slooten and Gill (FSI: Genetics, 2012) and implemented in e.g., LRMix Studio for this exercise. Assume the mixture is {1, 2, 3, 4}. Let LR_2 compare the hypothesis H_{p2} : *POI + unknown contributed* to H_{d2} : *two unknowns contributed*. Assuming that the drop out probability is $d = 0.05$ for all contributors, we find that LR_2 equals

- a. $(1-d)^2/(12p^2)$
- b. $1/(1-d)^2 * 1/(12p^2)$
- c. $d/(2p^2)$
- d. $1/(2dp^2)$
- e. $1/(12p^2)$

6) Consider question 1. Assume theta correction with $\theta = 0.02$ (and no drop-out, i.e., $d = 0$). In this case LR_1

- a. $1/(2p^2)$
- b. $(1 - \theta)/(2p^2)$
- c. $2(\theta + (1-\theta)p)/(2p^2)$

d.

$$\frac{(1 + \theta)(1 + 2\theta)}{2(\theta + (1 - \theta)p)(\theta + (1 - \theta)p)}$$

e.

$$\frac{2(\theta + (1 - \theta)p)(\theta + (1 - \theta)p)}{(1 + \theta)(1 + 2\theta)}$$

7) Assume $d = 0$ and $\theta = 0$. Consider 10 independent copies of the marker used above. The $\log_{10}(\text{LR})$ comparing the hypothesis H_{p1} : *POI contributed* to H_{d1} : *an unknown contributed* is

- a. 50^{10}
- b. 500
- c. 2.7
- d. 10
- e. 16.99

8) Assuming the hypothesis H_{P3} : *POI + two unknowns*, and taking into account the peak heights, we estimate the portion of the mixture that comes from POI to

- a. 0.426
- b. 0.414
- c. 0.840
- d. 0.420
- e. 1

Part B

The purpose of this part is to illustrate the top-down approach introduced in Slooten "A top-down approach to DNA mixtures", *Forensic Science International: Genetics* 46 (2020). The brief description below should suffice to do the exercises in this part. However, you need to read the mentioned paper to fully understand what is going on. Our goal is to find the LR comparing the hypothesis H_p : *POI contributed* to H_d : *an unknown contributed* without having to make strong, potentially dubious assumptions, typically needed for other models. It is for instance not necessary to specify how many contributors there are and we do not define a peak height distribution (EuroForMix assumes a gamma model, but other distributions like the log normal are used).

Initially, we only consider the evidence from the one marker analysed above, i.e.,

1 (426), 2 (414), 3(46), 4 (44), 5 (37), 6 (33)

The peak heights appear in descending order and their sum is $426 + \dots + 33 = 1000$. We define M_α as the sub profile of $M = \{1, 2, 3, 4, 5, 6\}$ that contains the smallest set of peaks such that the sum of the peak heights in M_α is at least a fraction α of the total sum of peak heights. On every locus, M_α can be iteratively constructed, by taking in peaks starting with the largest one, and stopping when a fraction α or more of the total sum of peak heights has been taken into M_α . We will use $\alpha = 0.1, 0.2, \dots, 1$. For instance, $M_{0.1} = \{1\}$ since $426/1000 = 0.426 > 0.1$. $LR_{0.1}$, is calculated with the minimum number of contributors needed to explain $M_{0.1} = \{1\}$, i.e., one contributor. The LR calculations are done using the simple discrete model and below we assume no drop-out, no drop-in, and $\theta = 0$. Since the POI has the genotype $1/2$, $LR_{0.1} = 0$. The final top-down likelihood ratio, $LR_{\text{top-down}}$, is the largest of $LR_{0.1}, \dots, LR_{0.9}, LR_{1.0}$.

9) We find

- $M_{0.1} = M_{0.2} = M_{0.3} = M_{0.4} = \{1\}$ and $LR_{0.1} = LR_{0.2} = LR_{0.3} = LR_{0.4} = 0$
- $M_{0.1} = M_{0.2} = M_{0.3} = M_{0.4} = \{1\}$ and $LR_{0.1} = LR_{0.2} = LR_{0.3} = 0, LR_{0.4} = 1/(2p^2)$
- $M_{0.1} = \{1\}, M_{0.2} = M_{0.3} = M_{0.4} = \{1, 2\}$ and $LR_{0.1} = 0, LR_{0.2} = LR_{0.3} = LR_{0.4} = 1$
- $M_{0.1} = \{1\}, M_{0.2} = M_{0.3} = M_{0.4} = \{1, 2\}$ and $LR_{0.1} = 0, LR_{0.2} = LR_{0.3} = LR_{0.4} = 1/(2p^2)$
- $M_{0.1} = \{1\}, M_{0.2} = M_{0.3} = M_{0.4} = \{1, 2\}$ and $LR_{0.1} = 0, LR_{0.2} = LR_{0.3} = LR_{0.4} = 1/p^2$

10) We find

- $LR_{0.5} = LR_{0.6} = LR_{0.7} = LR_{0.8} = 1/(12p^2)$
- $LR_{0.5} = LR_{0.6} = LR_{0.7} = LR_{0.8} = 1/(30p^2)$
- $LR_{0.5} = 0.5$
- $LR_{0.6} = 1$
- $LR_{0.5} = LR_{0.6} = LR_{0.7} = LR_{0.8} = 1/(2p^2)$

11) We find

- a. $LR_{0.9} = 0$
- b. $LR_{0.9} = 1/(12p^2)$
- c. $LR_{0.9} = 1$
- d. $LR_{0.9} = 1/(30p^2)$
- e. $LR_{0.9} = 1/(2p^2)$

12) We find

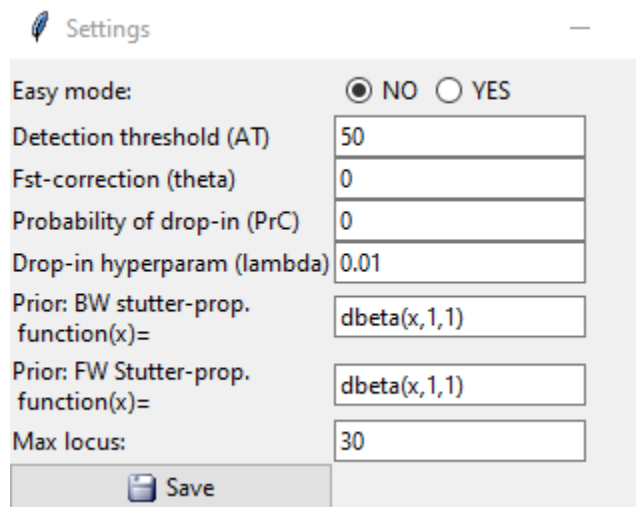
- a. $LR_{1.0} = 1/(30p^2)$
- b. $LR_{1.0} = 1$
- c. $LR_{1.0} = 0$
- d. $LR_{1.0} = 1/(2p^2)$
- e. $LR_{1.0} = 1/(12p^2)$

13) Based on the top down approach and 10 independent copies of the above marker we would report

- a. $\log_{10}(LR_{\text{top-down}}) = 2.7$
- b. $\log_{10}(LR_{\text{top-down}}) = 10$
- c. $\log_{10}(LR_{\text{top-down}}) = 50$
- d. $\log_{10}(LR_{\text{top-down}}) = 16.99$
- e. $\log_{10}(LR_{\text{top-down}}) = 500$

Part C Peak height data

For the problems below the freely available software EuroForMix, preferably version 3.3.1, available from <http://www.euroformix.com/>, is required. **You should use the settings:**



The screenshot shows the 'Settings' dialog box for EuroForMix. It contains the following settings:

- Easy mode: NO YES
- Detection threshold (AT): 50
- Fst-correction (theta): 0
- Probability of drop-in (PrC): 0
- Drop-in hyperparam (lambda): 0.01
- Prior: BW stutter-prop. function(x)=: dbeta(x,1,1)
- Prior: FW Stutter-prop. function(x)=: dbeta(x,1,1)
- Max locus: 30

At the bottom left, there is a 'Save' button with a floppy disk icon.

We do not consider stutter or drop-in, i.e. PrC=0, BW Stutter: NO, FW Stutter: NO (degradation is not activated for these markers as we are not using one of the predefined kits in EuroForMix). Use the maximum likelihood option ('Quantitative LR (Maximum Likelihood based)') in EuroForMix for LR calculations. Some theory is explained in [01-Introduccion-Quant.pdf](#) (by Lourdes Prieto) in case you are new to the software.

The evidence is [here](#). There are ten markers (L1, L2, ..., L10) of the kind described above (all with 10 alleles, each having frequency 0.1), in the file [freqs.csv](#). The POI has genotypes 1/2 for all markers. You need to prepare the input file for POI.

14) We compare the hypothesis Hp2: *POI and one unknown contributed* to Hd2: *two unknowns contributed*. EuroForMix reports

- a. $\log_{10}(\text{LR}) = 16.99$
- b. "Wrong model specification. The specified model could not explain the data. ..."
- c. $\log_{10}(\text{LR}) = 10.22$
- d. $\log_{10}(\text{LR}) = 11.53$
- e. $\log_{10}(\text{LR}) = 11.49$

15) If we compare Hp3: *POI and two unknowns contributed* to Hd3: *three unknowns contributed*, EuroForMix reports

- a. $\log_{10}(\text{LR}) = 10.22$
- b. "Wrong model specification. The specified model could not explain the data. ..."
- c. $\log_{10}(\text{LR}) = 11.53$
- d. $\log_{10}(\text{LR}) = 11.49$
- e. $\log_{10}(\text{LR}) = 16.99$

16) If we compare Hp1: *POI contributed* to Hd1: *one unknown contributed*, EuroForMix reports

- a. "Wrong model specification. The specified model could not explain the data. ..."
- b. $\log_{10}(\text{LR}) = 10.22$
- c. $\log_{10}(\text{LR}) = 11.53$
- d. $\log_{10}(\text{LR}) = 11.49$
- e. $\log_{10}(\text{LR}) = 16.99$

17) If we run 'Optimal quantitative LR' to estimate if 1, 2 or 3 contributors give the best model, EuroForMix

- a. reports "Wrong model specification"
- b. suggests a model with 2 contributors
- c. suggests a model with 1 contributor

- d. suggests a model with 3 contributors
- e. reports that the models fit equally well

For the remaining exercises, we consider the hypotheses Hp2: *POI and one unknown contributed* and Hd2: *two unknowns contributed*.

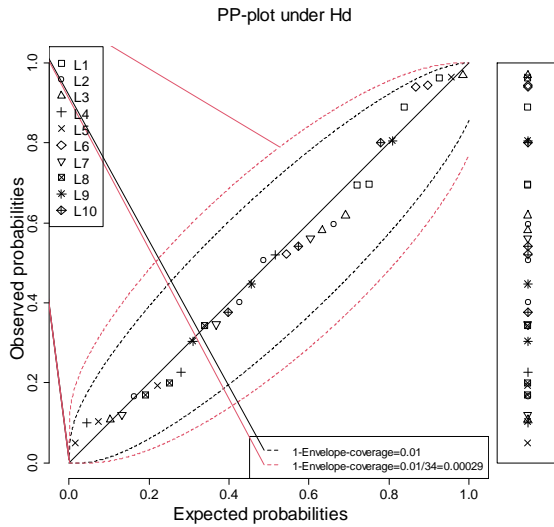
18) The predicted profile of the main contributor assuming Hd2

- a. has genotypes 1/2 for all markers and this is a reliable result
- b. has genotypes 1/1 for all markers and this is a reliable result
- c. has genotypes 1/2 for all markers but the result cannot be trusted
- d. cannot be inferred
- e. has genotypes 1/1 for all markers but the result cannot be trusted

19) The fraction contributed by the major contributor is estimated to

- a. 0.19
- b. 0.116
- c. 0.81
- d. 0.016
- e. 0.95

20) The plot



- a. indicates that there exists a better model than the one used
- b. indicates that the assumptions for calculations under Hd are violated
- c. does not indicate that the peak height assumptions for calculations under Hd are violated
- d. indicates that Hd is true
- e. indicates that Hd is false