Relationship Inference with Familias and R.
Statistical Methods in Forensic Genetics
Solutions all exercises

Thore Egeland
Daniel Kling
Petter Mostad

March 22, 2016

Chapter 2

Solutions: "Basics"

For some exercises, video solutions are also available as indicated by "Video available" in the heading of the solution. Links to these videos are available from `http://familias.name/VideosBook.pdf`.

Solution Exercise 2.1 (Video available)**.**

a) We can write

$$LR = \frac{\Pr(\text{child} \mid \text{mother,father})}{\Pr(\text{child} \mid \text{mother})}$$

Consider first the numerator. The only possible genotype for the child, given the mother and AF, is A/B, and therefore the probability is 1. For the denominator, the father must have passed on the A allele, and therefore the probability is p_A. Hence $LR = 1/p_A = 20$. The standard interpretation is "The data is 20 times more likely assuming AF to be the father compared to the alternative that some unknown man is the father".

b,c) See video.

d) $(1/p_A) * (1/p_a) = (1/0.05) * (1/0.1) = 200$

e,f) See video.

g) $RMP = p_A^2 p_a^2 = 1/40000$ and so $1/RMP = 40000$.

h) $LR = (1 + 3 \cdot 0.02)/(2 \cdot 0.02 + (1 - 0.02) \cdot 0.05) = 11.91$

i) Note that Hardy-Weinberg equilibrium is required for the LR derivation for each marker. This assumption is not needed when we use theta-correction. Furthermore, linkage equilibrium is needed. We have also assumed that there are no mutations or silent alleles. We assume AF and the mother to be unrelated.

Solution Exercise 2.2.

a) The numerator $\Pr(\text{data} \mid H_1) = 0$, and therefore also $LR = 0$.

b) $LR = 4.07e-03 = 0.00407$. This answer answer assumes that all alleles are entered; the answer difficult differs for all mutation models if only the alleles required are entered, in this case 14, 15, 16, 17 and a rest allele.

c) Note that $m = R/(n-1) = 0.007/7 = 0.001$. We find

$$LR = 0.001 \cdot (0.212 + 0.292)/(2 \cdot 0.212 \cdot 0.292) = 0.00407.$$

R is the expected mutation rate, m the probability for a specific mutation, for this model equal for all mutations.

Solution Exercise 2.3.

The pedigree corresponding to H_1 is specified as shown in Figure 2.1. Answers are obtained by loading the `Familias` file for this exercise. A table is obtained pressing `View Result` in the pedigree window, see Figure 2.2. The marker D7S820 gives a very large LR, namely 11189 since the allele 11.1 is so rare. If this marker is omitted the new LR is 530440484.5/11189.45872=47405.

Solution Exercise 2.4.

a) LR(grandfather/unrelated) = 0.98 for D3S1358.

b) LR(grandfather/unrelated) = 0.0085 for all markers.

c) Whether autosomal and haplotype markers can be combined and in case how is a big question and there appears to be no consensus.

Solution Exercise 2.5.

a) $W = 20/(1+20) = 0.952$.

b) $W = 200/(1+200) = 0.995$

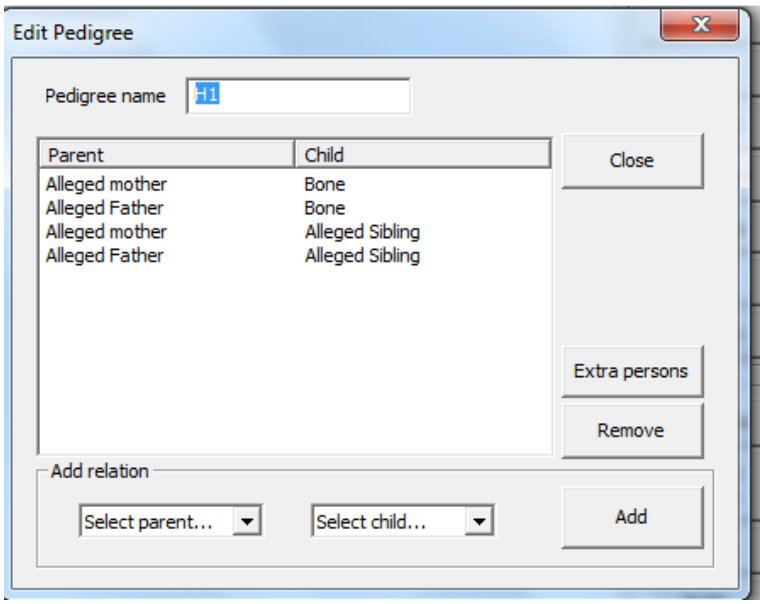


Figure 2.1: Defining the pedigree.

c) Same answers as above. Output for specific combinations of markers are conveniently found using `Included systems`.

d) This is a big question. Most recommendations favor LR. The problem with W is to decide on the prior. W may, however, be easier to interpret.

Solution Exercise 2.6.

a) We find $LR = 10$ and, using `Scale`, $LR(H_1/H_2) = 10, LR(H_1/H_3) = 1.905$.

b) The posteriors for H_1, H_2, H_3 are respectively $0.615, 0.0615, 0.323$.

Solution Exercise 2.7.

Most answers are given in the exercise and only a few additional comments are added. If a model is unstationary, allele frequencies change with generations. This adds some intuition as to why introducing an extra person, say a grandparent, typically changes results slightly for an unstationary model.

Consider the `Stepwise (unstationary)` model which gives $LR = 6.4e-03 = 0.0064$. Remove alleles so that only 14, 15, 16, 17 and a `Rest allele` remains. Then LR changes. This effect does not have to do with `Familias`;

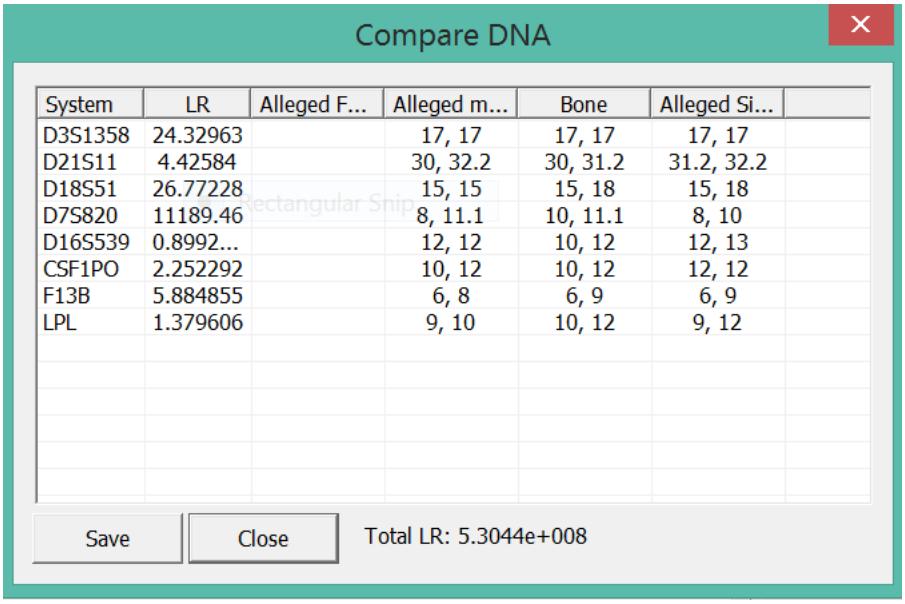


Figure 2.2: Output, `Familias` Exercise 2.3

rather the reason is that there are two different models, one with 8 alleles and one with fewer alleles (which constrains mutations within those alleles) and different models typically give different results. See Section 7.4 for a more complete discussion of mutation models.

Solution Exercise 2.8.

a) Assuming H_1, the child has inherited the allele 16 (17) from the mother with probability 1/2 and then one of the father's allele must have mutated to 17 (16). Therefore

$$\Pr(CH \mid AF) = \frac{1}{2}p_{16}(m_{14,17} + m_{15,17}) + \frac{1}{2}p_{17}(m_{14,16} + m_{15,16})$$

and the required result follows.

b) This follows by setting $m_{14,17} = m_{15,17} = m_{14,16} = m_{15,16} = m$. Then

$$LR = \frac{0.000714 \cdot (0.212 + 0.292)}{2 \cdot 0.212 \cdot 0.292} = 0.002907728.$$

c) We find

$$LR = \frac{p_{16} \cdot 2 \cdot kp_{17} + p_{17} \cdot 2 \cdot kp_{16}}{4p_{16}p_{17}} = k$$

by using the definition $m_{ij} = kp_j$. Furthermore,

$$R = \sum_{i=1}^{n} p_i \sum_{j \neq i} kp_j = \sum_{i=1}^{n} p_i k(1-p_i) \;\Rightarrow\; k = \frac{R}{\sum_{i=1}^{n} p_i(1-p_i)} = 0.00625953.$$

Solution Exercise 2.9. Regarding b): The marker with 0 LR, Penta_E is most easily found using `View result`. Regarding c), $LR = 4421152$, d) $LR(H_1/H_3) = 1.39$ (answers differ if mutations are only modelled for Penta_E). There is also solution file, `Solutions_2_9.fam` available. Regarding the last question, there is no consensus. One can argue that a model should be formulated before calculations and then appropriate mutation models should be specified for all markers. On the other hand, introducing mutations complicates calculations and this is a problem if it is desired to verify by hand. This is discussed at greater length in the Section 2.4.4 "Dealing with mutations in practice".

Solution Exercise 2.10.

a) $LR = 3.78$.

b) We could do simulations in `Familias`, see Exercise 2.17, and check if LR exceeds a specified threshold with an acceptable high probability. There are several ways to do the simulations, the most straightforward would be to load the database with the standard number of markers and simulate for these. X-chromosomal markers could also help and `FamLinkX` could then be used.

Solution Exercise 2.11.

The LR (father/not-father) is 1.363636. How to do it: Enter the allele system setting the silent allele frequency to 0.05. Enter the persons and their DNA data as usual. Construct the pedigrees manually and calculate.

Solution Exercise 2.12.

See Section 2.5.1 for an example based on the sampling formula for a similar case.

Solution Exercise 2.13.

$LR = 0.0068$.

Solution Exercise 2.14.

See video. The video is made for `Familias` 2, but the procedure is the same for `Familias` 3.

Solution Exercise 2.15.

a) LR (father/not father)=0 from `Familias`. The LR for marker PENTA_E is 0.

b) `General DNA data`: Click on the marker PENTA_E. In the new window click on options and set `Dropout` to 0.1. Save.
`Case-related DNA data`: choose the child and tick `Consider dropout` in the new window. A message will appear saying that `Familias` will model dropout, click OK. Calculate and find $LR = 2679875170$.

c) `Case-related DNA data`: Untick `consider dropout` for the child.
`General DNA data`: Choose mutation model for PENTA_E, see previous exercises on how to do this. Calculate $LR = 1078633$.

d) It's certainly not standard to use dropout routinely.

Solution Exercise 2.16.

Note that genotypes must be given as homozygotes in `Familias`.

a) `General DNA data`: when editing the allele data, choose `Options` and include a silent allele frequency of 0.05. Note that allele frequencies and the silent allele frequency should add to 1. Therefore, some change in allele frequencies may be required, for instance changing the rest allele frequency to 0.55. We find $LR = 0.57$.

b) `General DNA data`: Remove the silent allele frequency and include a dropout probability of 0.05. `Case-related DNA data`: Tick `Consider dropout` for both the alleged father and child. This gives LR=0.34.

Solution Exercise 2.17.

Simulation: In `Pedigrees` click `Simulate`. Move both AF and Child to `Will be genotyped`. The simulation will produce slightly different results each time it is run unless a seed is set. If you untick `random seed` and set seed to 12345, you should get the same results as below. Click `Simulate`. The mean LR is shown for both H_1 true and H_2 true.

a) The mean LR when H1 is true is 40.86.

b) The mean LR when H2 is true is 0.8979.

c) Click `LR limit`, choose LR threshold 50 and click update. The probability of observing a LR larger than 50 is 0.09.

Solution Exercise 2.18.

a) CSF1PO: The mother has transmitted allele 10, so the father must transmit allele 15. This happens with probability 1/2 under H_1 and with probability p_{15} under H_2. Thus, LR is $1/(2p_{15})$.

D7S820: Under H_1, the child's genotype has probability 1/2; under H_2 it has probability $\frac{1}{2}(p_{11} + p_{12})$), so $LR_{1,2} = 1/(p_{11} + p_{12})$.

D19S433: Under H_1 the child's genotype has probability 1; under H_2 it has probability p_{14}, so $LR_{1,2} = 1/p_{14}$.

b) CSF1PO: Now, the brother of the defendant must transmit allele 15. The allele he transmits is equal to allele 15 with probability 1/4, to allele 14 with probability 1/4, and is randomly drawn from the population with probability 1/2. Thus, he transmits allele 15 with probability $1/4+p_{15}/2$. Therefore

$$LR_{3,2} = \frac{\Pr(\text{child} \mid H_3)}{\Pr(\text{child} \mid H_2)} = \frac{\frac{1}{2}(\frac{1}{4} + \frac{1}{2}p_{15})}{\frac{1}{2}p_{15}} = \frac{1 + 2p_{15}}{4p_{15}}.$$

The answer can be checked using the file `Solutions2_18.fam`.

D7S820: The brother must transmit 11 or 12. This is done with probability $1/2+(p_{11}+p_{12})/2$. The mother's allele is the other one with probability 1/2, so probability of child's genotype under H_3 is $1/4 + (p_{11} + p_{12})/4$. Under H_2 we had probability $\frac{1}{2}(p_{11} + p_{12})$ so

$$LR_{3,2} = \frac{1 + p_{11} + p_{12}}{2(p_{11} + p_{12})}.$$

D19S433: The brother transmits allele 14 to the child with probability $\frac{1}{2}(1+p_{14})$ so $LR_{3,2} = (1+p_{14})/(2p_{14})$.

c) Yes: CSF1PO: Since $1/(2p_{15}) = 4.56$ we know p_{15} and this yields $LR_{3,2} = 2.78$.

D7S820: Since $1/(p_{11}+p_{12}) = 2.92$, we know $p_{11}+p_{12}$ and this yields $LR_{3,2} = 1.96$.

D19S433: Since $1/p_{14} = 2.93$, we know p_{14} and this yields $LR_{3,2} = 1.97$.

It may seem surprising that this is possible, but in the $LR_{1,2}$ only the matching allele(s) play a role. If there are two matching alleles such as 11 and 12 for D7S820, they may be viewed as a "11 or 12" allele.

d) Dividing yields $2/(1+2p_{15})$ for CSF1PO, $2/(1+p_{11}+p_{12})$ for D7S820 and $2/(1+p_{14})$ for D19S433.

e) The limit $p_{15} \to 1$ yields $LR_{1,3} = 2/3$ (support for the brother being the father), limit $p_{15} \to 0$ yields $LR_{1,3} = 1$. If allele 15 is very common, we would expect the brother to have more alleles 15 than the defendant and so have better chances to pass an allele 15 to an offspring. If, 15 is very rare, the evidence is neutral.

f) Now $LR_{1,3}$ is always greater than or equal to one, with equality if $p_{14} = 1$, in which case the brother must have two alleles 14 as well and we cannot distinguish them anymore.

g) Same kind of analysis. If $p_{11}+p_{12}$ is close to 1, $LR_{1,3} \approx 1$. If $p_{11}+p_{12}$ is close to 0, $LR_{1,3} \approx 2$ and there is evidence against the defendant.

h) No, since we do not have prior probabilities nor do we know if even more scenarios are possible (e.g., father of defendant is father of child?).

Chapter 3

Solutions: "Searching for relationships"

Solution Exercise 3.1 (Video available)**.**

a,b) The likelihood ratios are

$$LR_1 = 0,\ LR_2 = 8,\ LR_3 = 0,\ LR_4 = 1.$$

c,d) Scaling against unrelated is according to conventions. With flat priors

$$P(H_i \mid \text{data}) = \frac{L_i}{L_1 + L_2 + L_3 + L_4}.$$

and so

$$\begin{aligned} &P(H_1 \mid \text{data}) = 0, \\ &P(H_2 \mid \text{data}) = 8/9 = 0.89, \\ &P(H_3 \mid \text{data}) = 0, \\ &P(H_4 \mid \text{data}) = 1/9 = 0.11. \end{aligned}$$

e,f) The likelihood ratios are

$$LR_1 = 2,\ LR_2 = 0,\ LR_3 = 0,\ LR_4 = 1.$$

and the posterior becomes

$$P(H_1 \mid \text{data}) = 2/(2+1) = 0.667,$$
$$P(H_2 \mid \text{data}) = 0,$$
$$P(H_3 \mid \text{data}) = 0,$$
$$P(H_4 \mid \text{data}) = 1/(2+1) = 0.333.$$

See video for remaining questions.

Solution Exercise 3.2 (Video available)**.**

a)-d) See hints and video.

e) The LR for P1 versus P2 is 16 for a direct match and so the data is 16 times more likely for the hypothesis of a direct match compared to the unrelated hypothesis. Other results are interpreted similarly. P1 and P2 are therefore most likely the same person, they may also be siblings, $LR = 6.25$. Of course, this is a very simplified exercise using only one system. With several STR markers, two siblings will most probably not appear as a direct match due to the fact that they will most likely share no alleles (or only one) in some systems.

f) For the direct match the LR takes the form

$$LR = \frac{\Pr(\text{data} \mid P1 = P2 = 12/12)}{\Pr(\text{one is } 12/12 \mid \text{other is } 12/12)} = \frac{1}{p_{12}^2} = \frac{1}{0.25^2} = 16.$$

For the sibling match we find

$$LR = \frac{\Pr(\text{data} \mid \text{sibs})}{\Pr(\text{data} \mid \text{unrelated})} = \frac{\frac{1}{4}p_{12}^4 + \frac{1}{2}p_{12}^3 + \frac{1}{4}p_{12}^2}{p_{12}^4} = 6.25.$$

g) This will only affect the results for the direct match as `Familias` only accounts for these parameters in this case. Increasing the dropout probability will actually give slightly higher LR to the match between P1 and P2. Increasing the dropin probability will decrease the LR for the match between P1 and P2. Increasing only the typing error probability will produce other possible matches and will give a slightly lower LR for the direct match between P1 and P2.

Solution Exercise 3.3.

c) This may be a realistic scenario for different reasons. A simple reason may be that not all missing persons have been found. Another may be that not all remains produce DNA profiles. The fact that only 8 profiles is in the set even though the total number of missing persons is greater is accounted for in the prior.

d) This means we have some prior belief that the number of missing persons is 10.

l) For results see Figure 3.1.

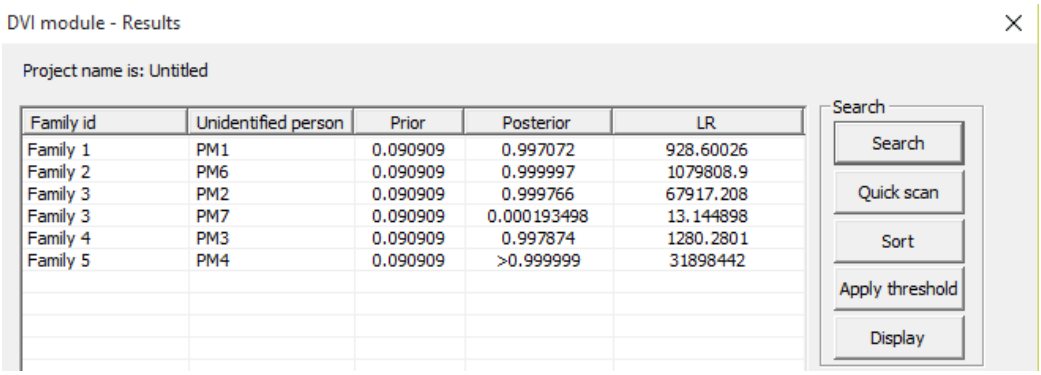


Figure 3.1: List of results from the DVI search for Exercise 3.3 l)

m) Not all remains where identified, this is expected as we only have reference data from 5 families. All posterior probabilities are above 99% (except for the match between Family 3 and PM7) though only three are greater than 99.99%.

n) The user will find a possible mutation for the match between Family 4 and PM3 for the marker vWA.

o) For results see Figure 3.2. We see that PM7 and PM8 has a possible sibling relation. If either of the two persons match in a family we may use this information to match both into that family. We may also combine this information with meta data such as known relationship between missing persons.

p) The posterior becomes considerably lower as the priors are lowered.

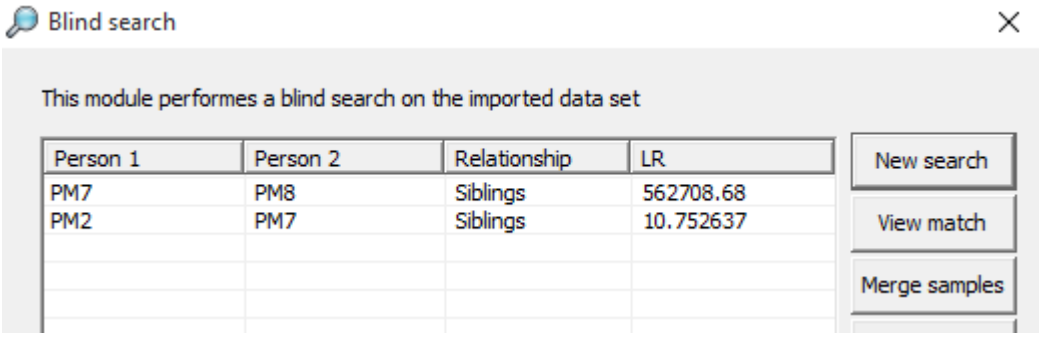


Figure 3.2: Results for Exercise 3.3 o).

q) * One way is to add another pedigree in the reference family. Another solution may be to add another reference family with the same reference person. The difference would be how the posteriors are calculated. We will use the first option, i.e., add another pedigree to Family 1, where we now need to define extra persons in order to define the brother relationship.

r) * For results see Figure 3.3. We see that we now have 3 possible matches for Family 1b and one for Family 1. We see that PM5 has the highest LR.

s) A better solution, but more complex, would be to allow the definition of several missing persons in the same pedigree. Familias would then either search for each missing persons individually, or try matching all unidentified persons with the missing persons at once. The complexity using the latter approach grows exponentially with the number of missing persons.

Solution Exercise 3.4.

b) * For results see Figure 3.4.

b) We see the same matches as in the previous exercise with slightly different LRs. This is partly due to the fact that the Quick scan feature

Project name is: Untitled Number of matches: 9

Family id	Unidentified person	Prior	Posterior	LR	Systems used	#Mismatches
Family 1	PM1	0.090909	0.99678	928.60026	15	0
Family 2	PM6	0.090909	0.999997	1079808.9	15	0
Family 3	PM2	0.090909	0.999762	67917.208	15	0
Family 3	PM7	0.090909	0.000193497	13.144898	15	2
Family 4	PM3	0.090909	0.997662	1280.2801	15	1
Family 5	PM4	0.090909	>0.999999	31898442	15	0
Family 1b	PM1	0.090909	0.00082709	7.478567	15	0
Family 1b	PM4	0.090909	0.000260125	2.3520553	15	0
Family 1b	PM5	0.090909	0.998581	9029.1968	15	0

Figure 3.3: Results for Exercise 3.3 r).

assumes a zero mutation rate model for all relationships except parent-child.

c) A quick scan is extremely fast using no information about complex pedigree structures. Also, sometimes relationships may be erroneously specified or unspecified. The quick scan will perform a swift search before doing the complete search.

h) A minor change indicating that there is a low probability that PM5 is the child (or parent) of the individual in Family 1. The LR is only 0.022. Examining the match, we find two possible mutations necessary for this relation to be true.

i) There is no "perfect" value for the number of mismatches, but three (3) should be an "acceptable" value. This specifies that we discard any matches where the number of mismatches exceeds 3. For all other matches we compute a LR, in other words if we encounter three possible mutations for a parent child relation, we still compute an LR. Setting the value to0 low, i.e., 0 or 1, may cause true matches to be missed, in other words an inflated false negative rate.

Solution Exercise 3.5.

c) *Comment*: See manual at `http://familias.no/english/manual/` for a comprehensive list of relationships that `Familias` recognizes.

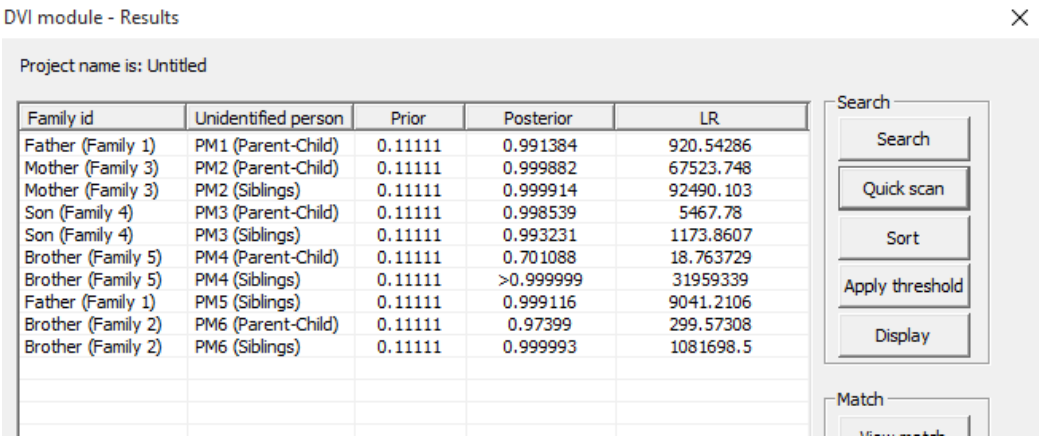


Figure 3.4: Results for Exercise 3.4 a).

i) For results see Figure 3.5.

j) We find some spurious matches, which can always expected as siblings may share zero alleles for all typed markers, however with a low probability. Typically such relationships are resolved using more markers.

[l) A hint is to sort using Family id which will make it easier to distinguish any matches that has been falsely reported or missed. Sister20 is missing, performing a search with a match threshold of 1 reveals that this match has a LR of only 9. In addition, Sister64 is missing from the list, similar investigation as described previously reveals a LR of only 0.99 for this match. We further find that there are a number of false matches (exceeding the threshold of 10). A hint is to sort by LR and investigate the lower matches.

Solution Exercise 3.6.

Answers are given in the exercise.

Solution Exercise 3.7.

c) The number of comparisons are found using the arithmetic series

$$999 + 998 + \cdots + 1 = \frac{999(999+1)}{2} = 499,500 \tag{3.1}$$

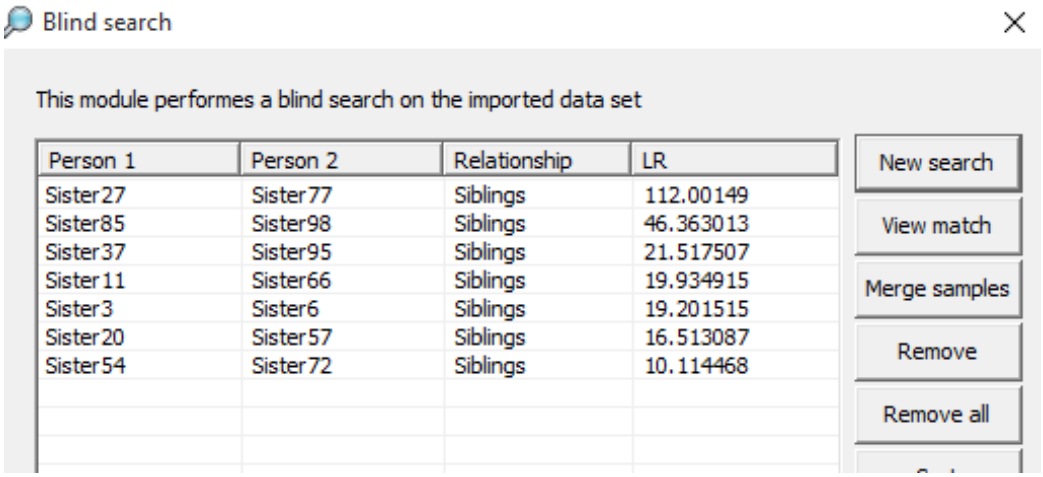


Figure 3.5: Results for Exercise 3.5 i).

In other words, nearly half a million comparisons are performed. We also need to keep in mind that for each comparison we actually perform a number of computations, one for each overlapping system.

d) There should be one possible match between sample 421 and 501 with a LR of 9.8. Setting appropriate values on those parameters is extremely difficult and depends to a large degree on the quality of the sample. The default values are probably good enough to generally account for low quality samples.

e) Using the same reasoning as in c) we can illustrate this with an arithmetic series and get the total number of comparisons as 12,497,500,000. In other words, more than 12 billion comparisons would be performed. The blind search function in `Familias` would probably not cope with the large number of comparisons. Other means, e.g., external software, would be needed to certify that the database does not contain duplicate entries.

h) For results see Figure 3.6.

We have four distinct matches. Looking closer at each of the matches, it seems that none of them are really a direct match, the results suggest another relationship, e.g., siblings?

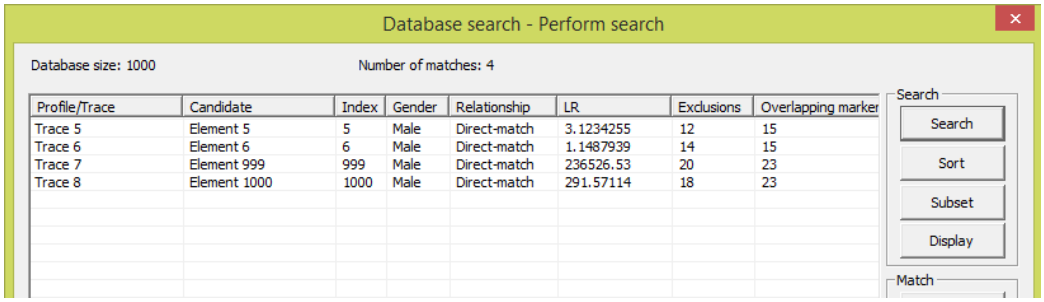


Figure 3.6: Results for Exercise 3.7 h).

i) See Figure 3.7 for the top results.

j) For results see Figure 3.8. Only the top-10 matches are kept.

k) For results see Figure 3.9. Only matches with an LR exceeding 100 are kept. In total 19 matches are kept.

l) Using the top-k method we are left with a specified number of matches while with the LR-threshold method we may possibly end up with a large amount of matches. The top-k method may, on the other hand, miss matches where the LR is high.

m) For results see Figure 3.10. We are left with 6 matches.

n) For results see Figure 3.11, where we have used alpha=0.5. The number of matches that are kept is increased, alpha is the true positive rate used in the conditional simulations. Conditional on the profile of interest, we compute the conditional LR distribution and based on this we find the LR threshold at the point where 0.5 (alpha) of the LR is above.

o) These settings may be relevant when we have a database from a subpopulation or when there is reason to believe that the database is inbred.

p) For results see Figure 3.12.

Solution Exercise 3.8.

b) The formulae below gives $LR = 6.25$:

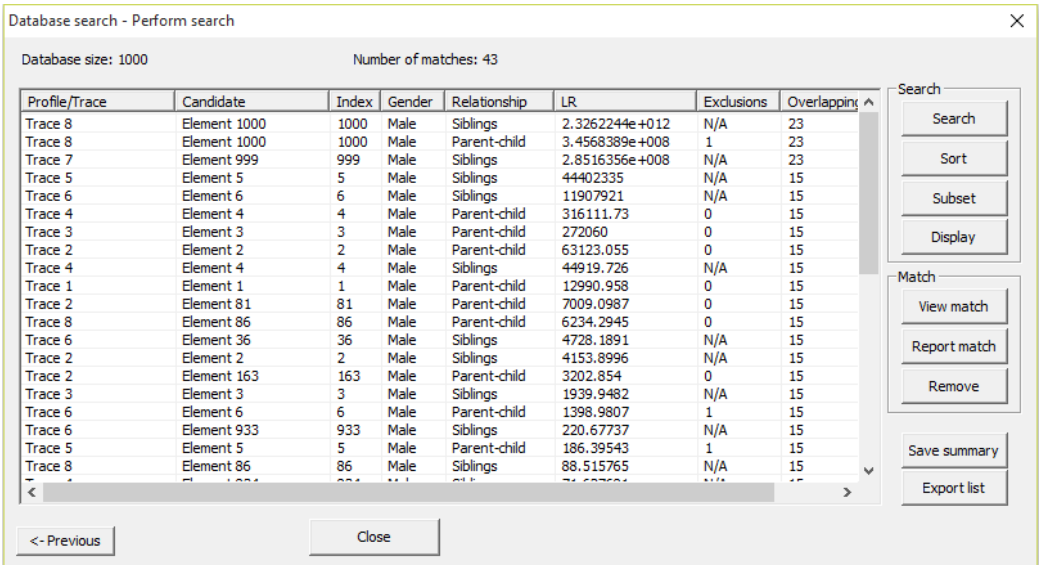


Figure 3.7: Results for Exercise 3.7 i).

```
p12 <- 0.1; p13 <- 0.2; p14 <- 0.3; p15 <- 0.2; p16 <- 0.1; p17 <- 0.1
L1 <- p14^2+2*p14*(p12+p13)
t1 <- (p12^2*2*p13*p14+p13^2*2*p12*p14+p14^2*2*p12*p13)*2
t2 <- 2*4*p12*p13*p14*(p12+p13+p14)
L2 <- t1+t2
(LR <- L1/L2)
```

c) For results see Figure 3.13.

d) $LR = 2.291667$.

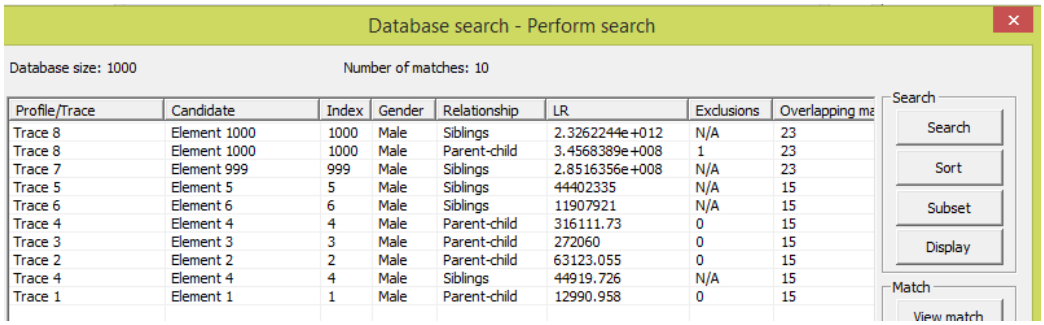
e) For results see Figure 3.14.

f) $LR = 2.5$.

g) $LR = 2.958333$

h) For results see Figure 3.15.

i) We can expect a high LR for the match between P_1 and S_2 as two alleles are shared whereas for the other two profiles (S_3 and S_4) lower LRs are excepted.



Profile/Trace	Candidate	Index	Gender	Relationship	LR	Exclusions	Overlapping ma
Trace 8	Element 1000	1000	Male	Siblings	2.3262244e+012	N/A	23
Trace 8	Element 1000	1000	Male	Parent-child	3.4568389e+008	1	23
Trace 7	Element 999	999	Male	Siblings	2.8516356e+008	N/A	23
Trace 5	Element 5	5	Male	Siblings	44402335	N/A	15
Trace 6	Element 6	6	Male	Siblings	11907921	N/A	15
Trace 4	Element 4	4	Male	Parent-child	316111.73	0	15
Trace 3	Element 3	3	Male	Parent-child	272060	0	15
Trace 2	Element 2	2	Male	Parent-child	63123.055	0	15
Trace 4	Element 4	4	Male	Siblings	44919.726	N/A	15
Trace 1	Element 1	1	Male	Parent-child	12990.958	0	15

Figure 3.8: Results for Exercise 3.7 j).

j) For results see Figure 3.16.

Solution Exercise 3.9.

e) For results see Figure 3.17.

g) The statistics file is available in the online repository at `http://familias.name`, see solution files.

Solution Exercise 3.10.

d) For results see Figure 3.18. The LR is computed as $1/(2p_{18}p_{19})$

e) For results see Figure 3.19.

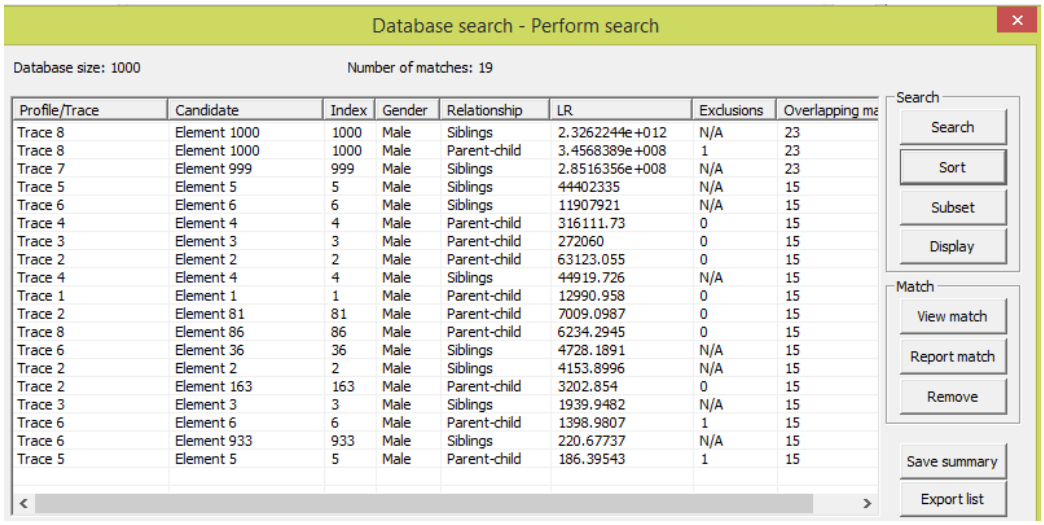
g) We find $LR = d(1-d)^3/p_{13}^2 = 1008.997$.

h) The reason for the high LR is the combination of a comparatively high dropout probability and the rarity of the shared allele (13). As for the match between P9 and P10, the explanation for the decreased LR is due the fact that we consider dropouts. As a consequence the true genotype may be a different from the one observed and we have instead a summation over different possible genotypes.

i) For results see Figure 3.20.

j) We find $LR = c(1-c)/(2p_{13}) = 5.294118$.

k) For results see Figure 3.21.

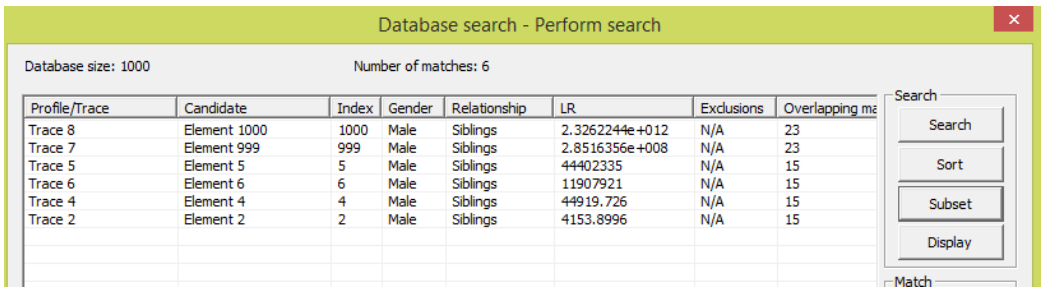


Profile/Trace	Candidate	Index	Gender	Relationship	LR	Exclusions	Overlapping ma
Trace 8	Element 1000	1000	Male	Siblings	2.3262244e+012	N/A	23
Trace 8	Element 1000	1000	Male	Parent-child	3.4568389e+008	1	23
Trace 7	Element 999	999	Male	Siblings	2.8516356e+008	N/A	23
Trace 5	Element 5	5	Male	Siblings	44402335	N/A	15
Trace 6	Element 6	6	Male	Siblings	11907921	N/A	15
Trace 4	Element 4	4	Male	Parent-child	316111.73	0	15
Trace 3	Element 3	3	Male	Parent-child	272060	0	15
Trace 2	Element 2	2	Male	Parent-child	63123.055	0	15
Trace 4	Element 4	4	Male	Siblings	44919.726	N/A	15
Trace 1	Element 1	1	Male	Parent-child	12990.958	0	15
Trace 2	Element 81	81	Male	Parent-child	7009.0987	0	15
Trace 8	Element 86	86	Male	Parent-child	6234.2945	0	15
Trace 6	Element 36	36	Male	Siblings	4728.1891	N/A	15
Trace 2	Element 2	2	Male	Siblings	4153.8996	N/A	15
Trace 2	Element 163	163	Male	Parent-child	3202.854	0	15
Trace 3	Element 3	3	Male	Siblings	1939.9482	N/A	15
Trace 6	Element 6	6	Male	Parent-child	1398.9807	1	15
Trace 6	Element 933	933	Male	Siblings	220.67737	N/A	15
Trace 5	Element 5	5	Male	Parent-child	186.39543	1	15

Figure 3.9: Results for Exercise 3.7 k).

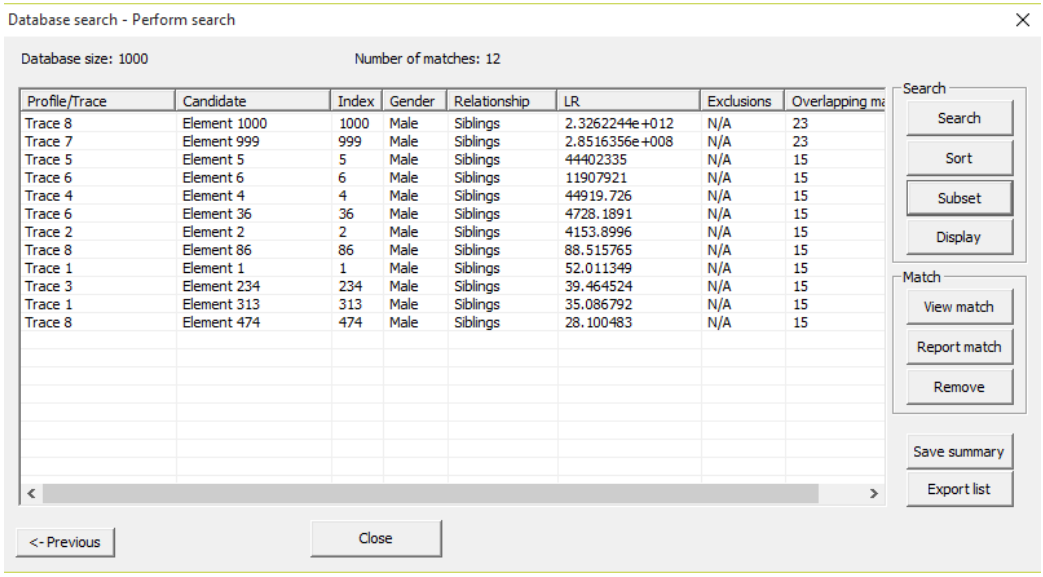
l) We see in Figure 3.21 that some cases give a lower LR while others lead to a higher LR. The explanation is that given the likelihood given that the two profiles are siblings yields a higher or lower likelihood than what does the likelihood given that the two profiles are unrelated. See derivation in m) for an example.

m) We find $LR = 4d(1-d)^3/(p_{13}^2 + p_{13}) = 34.01$.



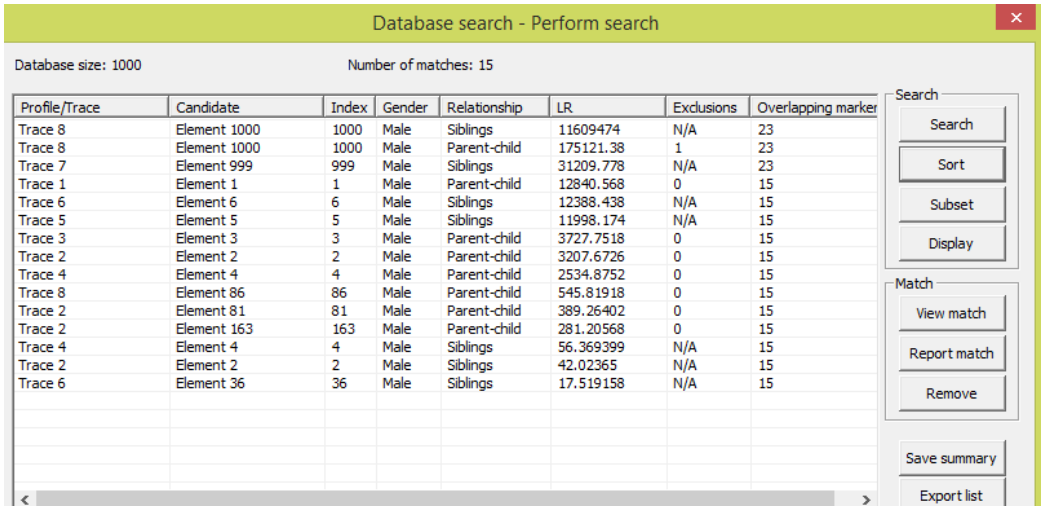
Profile/Trace	Candidate	Index	Gender	Relationship	LR	Exclusions	Overlapping ma
Trace 8	Element 1000	1000	Male	Siblings	2.3262244e+012	N/A	23
Trace 7	Element 999	999	Male	Siblings	2.8516356e+008	N/A	23
Trace 5	Element 5	5	Male	Siblings	44402335	N/A	15
Trace 6	Element 6	6	Male	Siblings	11907921	N/A	15
Trace 4	Element 4	4	Male	Siblings	44919.726	N/A	15
Trace 2	Element 2	2	Male	Siblings	4153.8996	N/A	15

Figure 3.10: Results for Exercise 3.7 m).



Profile/Trace	Candidate	Index	Gender	Relationship	LR	Exclusions	Overlapping ma
Trace 8	Element 1000	1000	Male	Siblings	2.3262244e+012	N/A	23
Trace 7	Element 999	999	Male	Siblings	2.8516356e+008	N/A	23
Trace 5	Element 5	5	Male	Siblings	44402335	N/A	15
Trace 6	Element 6	6	Male	Siblings	11907921	N/A	15
Trace 4	Element 4	4	Male	Siblings	44919.726	N/A	15
Trace 6	Element 36	36	Male	Siblings	4728.1891	N/A	15
Trace 2	Element 2	2	Male	Siblings	4153.8996	N/A	15
Trace 8	Element 86	86	Male	Siblings	88.515765	N/A	15
Trace 1	Element 1	1	Male	Siblings	52.011349	N/A	15
Trace 3	Element 234	234	Male	Siblings	39.464524	N/A	15
Trace 1	Element 313	313	Male	Siblings	35.086792	N/A	15
Trace 8	Element 474	474	Male	Siblings	28.100483	N/A	15

Figure 3.11: Results for Exercise 3.7 n).



Profile/Trace	Candidate	Index	Gender	Relationship	LR	Exclusions	Overlapping marker
Trace 8	Element 1000	1000	Male	Siblings	11609474	N/A	23
Trace 8	Element 1000	1000	Male	Parent-child	175121.38	1	23
Trace 7	Element 999	999	Male	Siblings	31209.778	N/A	23
Trace 1	Element 1	1	Male	Parent-child	12840.568	0	15
Trace 6	Element 6	6	Male	Siblings	12388.438	N/A	15
Trace 5	Element 5	5	Male	Siblings	11998.174	N/A	15
Trace 3	Element 3	3	Male	Parent-child	3727.7518	0	15
Trace 2	Element 2	2	Male	Parent-child	3207.6726	0	15
Trace 4	Element 4	4	Male	Parent-child	2534.8752	0	15
Trace 8	Element 86	86	Male	Parent-child	545.81918	0	15
Trace 2	Element 81	81	Male	Parent-child	389.26402	0	15
Trace 2	Element 163	163	Male	Parent-child	281.20568	0	15
Trace 4	Element 4	4	Male	Siblings	56.369399	N/A	15
Trace 2	Element 2	2	Male	Siblings	42.02365	N/A	15
Trace 6	Element 36	36	Male	Siblings	17.519158	N/A	15

Figure 3.12: Results for Exercise 3.7 p).

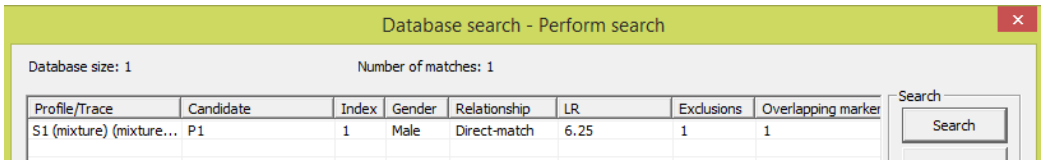


Figure 3.13: Results for Exercise 3.8 b).

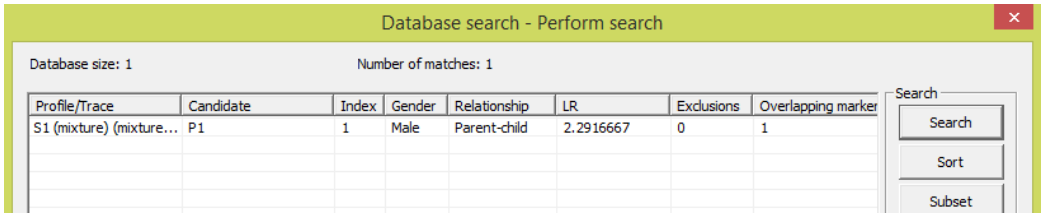


Figure 3.14: Results for Exercise 3.8 e).

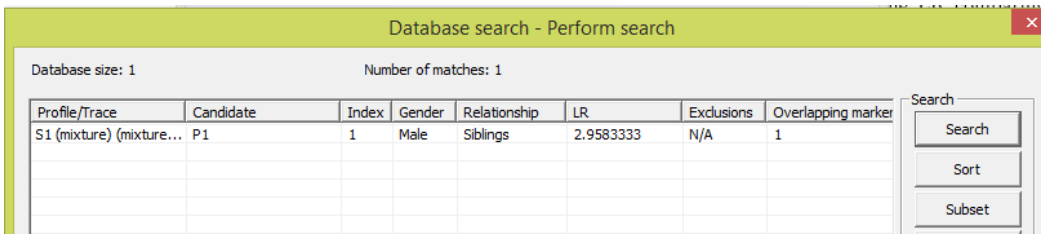


Figure 3.15: Results for Exercise 3.8 h).

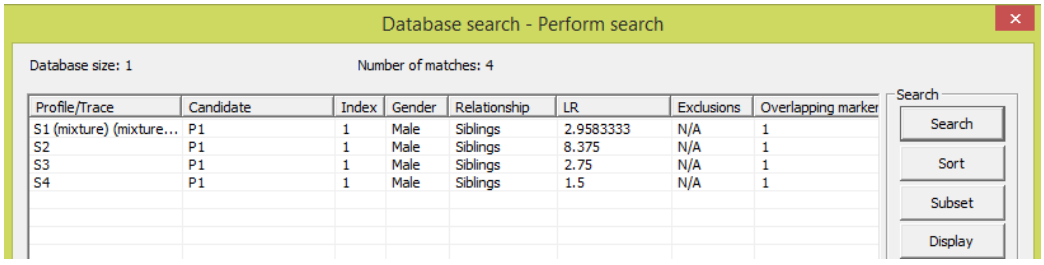


Figure 3.16: Results for Exercise 3.8 j).

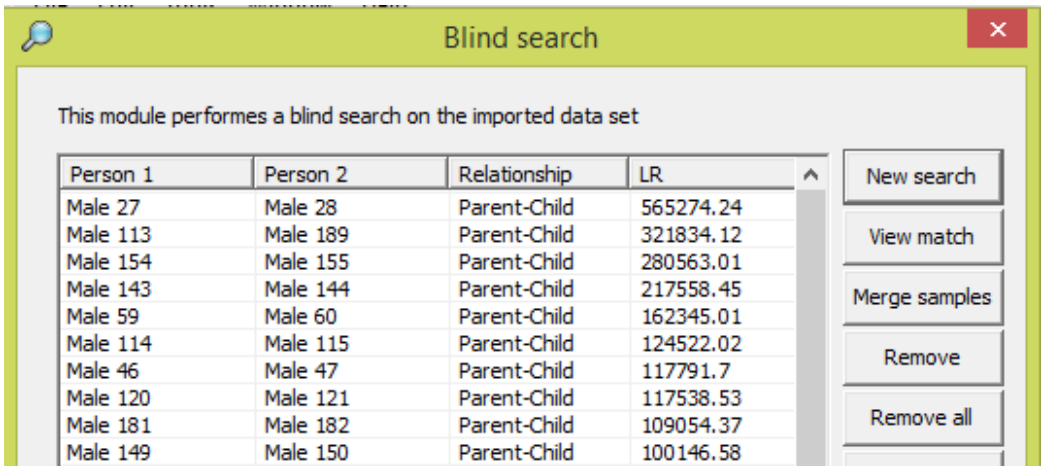


Figure 3.17: Results for Exercise 3.9 e).

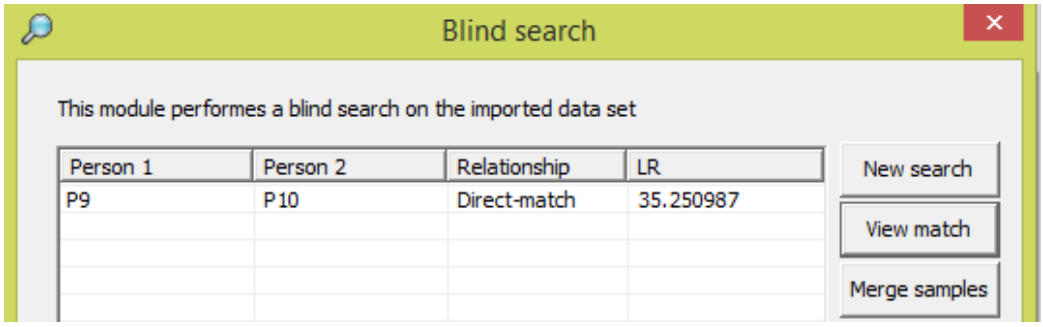


Figure 3.18: Results for Exercise 3.10 d).

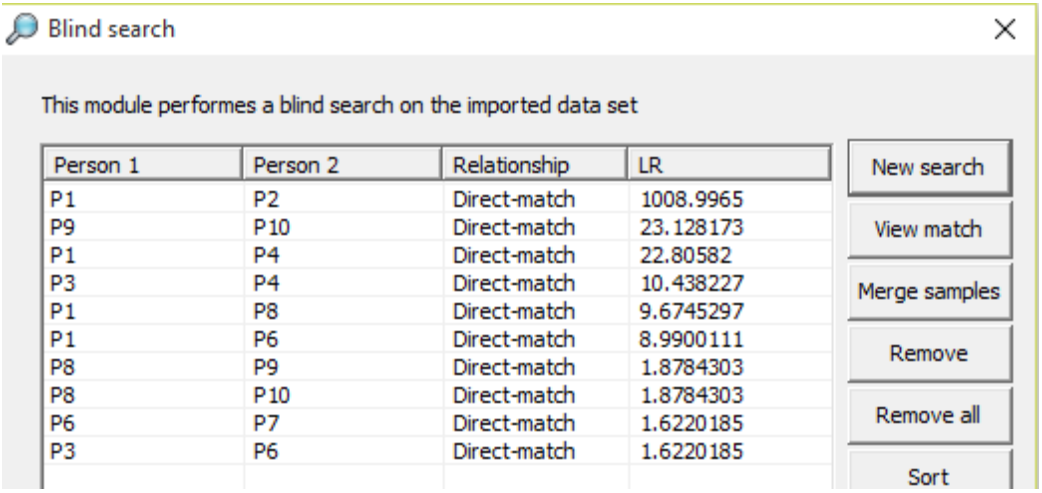


Figure 3.19: Results for Exercise 3.10 e).

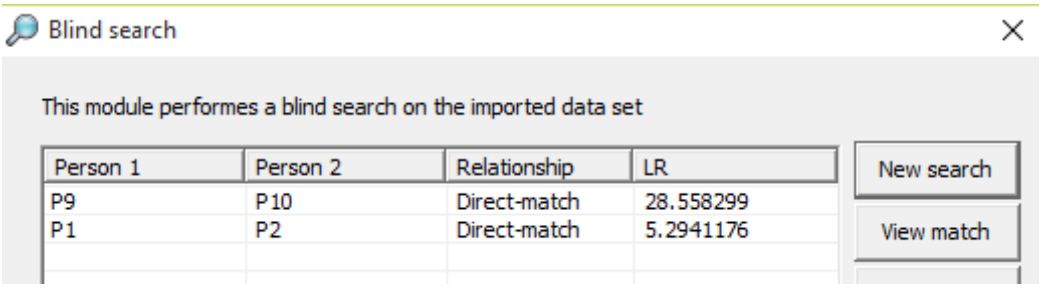


Figure 3.20: Results for Exercise 3.10 i).

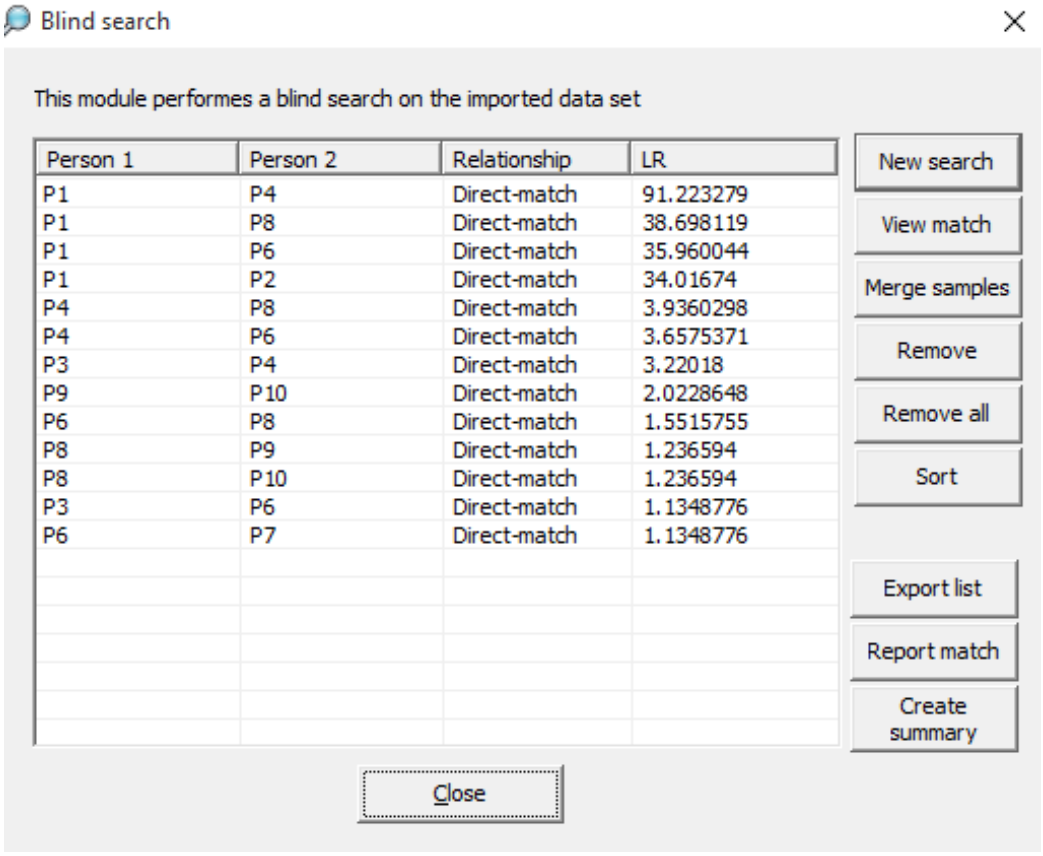


Figure 3.21: Results for Exercise 3.10 k).

Chapter 4

Solutions: "Dependent markers"

4.1 Autosomal markers and FamLink

Solution Exercise 4.1 (Video available)**.**

a) See video. *Comment*: Recall that a centiMorgan is a unit for measuring genetic linkage and that, as a rule of thumb, 1cM = 1% recombination rate. We use Haldane's or Kosambi's mapping functions, the former is the more common, to relate the recombination rate (r) to the genetic distance (d) in centiMorgans. `FamLink` implements Haldane's function

$$r = \frac{1 - \exp^{-2d/100}}{2}$$

In `FamLink` the above is accomplished by `Tools -> cM ...`.

b) The posterior should be 0.961 in favor of paternity.

c) The LR will remain the same regardless of the choice of the recombination rate. This can be shown mathematically, see e), while the intuitive answer is that there is not information on phase from a trio.

d) We may formulate the LR as

$$LR = \frac{1}{2p_{12}} \cdot \frac{1}{2p_{12}} = \frac{1}{4 \cdot 0.1^2} = 25$$

given that the two markers are unlinked and in linkage disequilibrium. To prove that the same applies also for linked markers it is in *this* case sufficient to note that as the father is homozygous for both markers, recombination has no impact.

e) The LR is 1.952 accounting for linkage and 2.777 when not accounting for linkage

f) `File -> Save`.

Solution Exercise 4.2.

a) See hints.

b) The LRs when accounting for linkage is LR(Full siblings)=1924.2 and LR(Half siblings)=43.9. It is reasonable that the LR for the full siblings alternative should be larger than for the half sibling case The corresponding posterior probabilities become W(Full siblings)=0.977 and W(Half siblings)=0.022.

c) The requested LR comparing full siblings and half siblings can be derived as, LR(Full siblings)/LR(Half siblings) from a), and the resulting ratio is $43.9 \approx 44$. In other words, given the data, it is 44 times more probable that P1 and P2 are full siblings rather than half siblings.

d) The LR(linkage) and LR(no linkage) will now coincide as a recombination of 0.5 implies independence.

e) No, autosomal markers cannot be used to distinguish paternal from maternal half siblings as the inheritance patterns are identical. A slightly different LR can be obtained given we have different mutation models for female and male transitions.

Solution Exercise 4.3.

a) The LR comparing H_1 versus H_3, $LR_{1,3}$, is 11.3 while the LR comparing H_2 versus H_3, $LR_{2,3}$, is 12.2. Even though the difference given the current data is fairly small, the results illustrate that linked markers can distinguish alternatives that are symmetric using unlinked markers. (The latter is indicated in the LR(no linkage) results). The LR comparing H_1 versus H_2, $LR_{1,2}$, is 0.926, suggesting that the data is slightly more likely given that H_2 is true.

b,c) The answer is given in the previous discussion, i.e., the LR will change to 9.0 for both $LR_{1,3}$ and $LR_{2,3}$.

d) $LR_{1,3}$ now becomes 9.75 while $LR_{2,3} = 11$ and $LR_{1,2} = 0.886$.

Solution Exercise 4.4.

d) The LR accounting for linkage is 65.1 while the LR when not accounting for linkage is 42.25. This suggests that the evidence is slightly underestimated when not accounting for linkage.

e) $L(H_P)$ is calculated as $p_{14} \cdot p_{21} = 0.3 \cdot 0.2 = 0.06$. $L(H_{D2})$ is calculated as $p_{14}^2 \cdot p_{21}^2 = 0.3^2 \cdot 0.2^2 = 0.0036$. The LR is computed as $0.06/0.0036 = 277$

f) The LR is computed by dividing the LR comparing H_P to H_{D2} with the LR comparing H_{D1} to H_{D2}. The LR we seek is $277/65.1 = 4.25$

f) The multiplication factor is the value comparing LR(linkage) with LR(no linkage), in the current case given by the ratio 65.1/42.25. It gives an indication of the over/underestimation of the LR when not accounting for linkage. The value can be combined with the LR for calculations for multiple other markers.

h) We compare H_P with the new defence hypotheses, similarly as in f), and get LR=277/7.9=35 for uncle and LR=277/8.4=33 for grandparent.

Solution Exercise 4.5.

d) Exploring the contents of the simulation reort we notice that if we simulate H_2 we will have situations where the data given H_1 indicates genetic inconsistencies. As `FamLink` does not model mutations, the simulations given H_2 are not as relevant.

e) 0.632. *Comment*: to find the median effect of linkage, we must consider the table comparing LR(no linkage) with LR(linkage).

f) 1.14 (Hint: in `Excel`, select cell P1 and insert "=AVERAGE(P6:P1007)", where AVERAGE may be substituted to the proper word in your language.)

g) Assuming we simulate data given H_2, genetic inconsistencies can be observed under H_1 thus yielding LR=0. The exact number is found in the simulation report.

h) As the number of simulations where the LR is zero is high (907) we need a greater number of simulations. This is also generally true. Also, the necessary LR is not indicated in the raw data output. We must calculate 1/LR presented in column D. Taking the average over the results should provide a value closer to 1, given that the number of simulations approaches infinity. In other words, running say 100,000 simulations we can expect the average to be closer to 1.

Solution Exercise 4.6.

a) The recombination rate is approximately 0.042, calculated using Haldane's mapping function. For small values of genetic distance the recombination rate is approximately equal to $d/100$, where d is the distance in cM.

d) The LR is 0.75 when accounting for linkage and 2.055 when not accounting for linkage. The multiplication factor is calculated as $0.75/2.055 = 0.36$, i.e., the LR is overestimated by a factor of more than 2 when not accounting for linkage.

e) From other software, not accounting for linkage, we already have LR(no linkage) for the two markers, e.g., from Familias. The multiplication factor is the ratio LR(linkage)/LR(no linkage). By multiplying the overall LR (obtained in another software) with the multiplication factor, the LR(no linkage) for the two markers cancels out thus obtaining a new overall LR where linkage is accounted for.

f) The median is given by 0.59 while the percentiles are 5%=0.41 and 95%=4.0 when H_1 is true. When H_2 is true, the median is given by 0.54 with percentiles is 5%=0.54 and 95%=3.16

Solution Exercise 4.7.

d) The `Search and subtract` method will make sure that the total frequencies sum to one once the calculations are performed by searching the frequency database and putting all alleles not observed in the current case into one rest allele. If the total sum of allele frequencies for any given system is above 1, the method will remove some frequency probability distribution from the rest allele. This method only works when not accounting for mutations. Unless the simple mutation model is considered, each allele will have different weights and thus we must treat

them separately. In `Familias`, the method would only be applicable if we do not consider mutations or if the `Equal probability (Simple)` is used.

e) The LR(linkage) is 3979202 while the LR(no linkage) is 996836; a large LR in favor of H_1.

f) Alternative hypotheses should be considered as hypotheses where one of the siblings are a half sibling or unrelated while the other two are full siblings. As H_2 now indicates three unrelated individuals and H_1 indicates full siblings, a high LR can be expected even though one of the individuals is a half sibling to the others.

Solution Exercise 4.8.

c) The LR(linkage) is 2401 while the LR(no linkage) is 1123. The corresponding multiplication factor is 2.1. The new combined LR becomes $500 \cdot 2.1 = 1050$. This may very well change our conclusion, provided the threshold for providing a positive conclusion, i.e., stating that the data favors H_1, is 1000.

Solution Exercise 4.9.

a) One solution is given below.

We first derive the likelihood given H_1 as

$$\begin{aligned}
L(H_1) &= \Pr(\text{data} \mid H_1) \\
&= \Pr(I_{L1}=0)\Pr(\text{data} \mid I_{L1}=0)\big(\Pr(I_{L2}=0 \mid I_{L1}=0)\Pr(\text{data} \mid I_{L2}=0) \\
&+ \Pr(I_{L2}=1 \mid I_{L1}=0)\Pr(\text{data} \mid I_{L2}=1)\big) \\
&+ \Pr(I_{L1}=1)\Pr(\text{data} \mid I_{L1}=1)\big(\Pr(I_{L2}=0 \mid I_{L1}=1)\Pr(\text{data} \mid I_{L2}=0) \\
&+ \Pr(I_{L2}=1 \mid I_{L1}=1)\Pr(\text{data} \mid I_{L2}=1)\big) \\
&= 0.5 \cdot 4p_9p_{12}^2p_{15}\Big(\big((1-r)^2+r^2\big)4p_{19}p_{21}^2p_{25} + 2r(1-r)p_{19}p_{21}p_{25}\Big) \\
&+ 0.5 \cdot p_9p_{12}p_{15}\Big(2r(1-r)4p_{19}p_{21}^2p_{25} + ((1-r)^2+r^2)p_{19}p_{21}p_{25}\Big)
\end{aligned}$$

$$
\begin{aligned}
&L(H_2) = \Pr(\text{data} \mid H_2) \\
&= 1/32 \cdot 4p_9p_{12}^2p_{15}\Big(\big((16(1-r)^5 + 40(1-r)^4r + 64(1-r)^3r^2 + 80(1-r)^2r^3 + 48(1-r)r^4 + 8r^5)\big)4p_{19}p_{21}^2p_{25} \\
&\qquad + \big(40(1-r)^4r + 96(1-r)^3r^2 + 80(1-r)^2r^3 + 32(1-r)r^4 + 8r^5\big)p_{19}p_{21}p_{25}\Big) \\
&+1/32 \cdot p_9p_{12}p_{15}\Big(\big(40(1-r)^4r + 96(1-r)^3r^2 + 80(1-r)^2r^3 + 32(1-r)^4 + 8r^5\big)4p_{19}p_{21}^2p_{25} \\
&\qquad + \big(16(1-r)^5 + 40(1-r)^4r + 64(1-r)^3r^2 + 80(1-r)^2r^3 + 48(1-r)r^4 + 8r^5\big)p_{19}p_{21}p_{25}\Big)
\end{aligned}
$$

We omit details on how to finally derive the LR as this is easily done by simply dividing $L(H_1)$ with $L(H_2)$.

b) * The LR becomes 1.015. The exact LR is 1.0157..., use the export LR function to get more decimals.

c) See Figure 4.1.

Solution Exercise 4.10.

b) See results in Figure 4.2.

e) The LR computed in `FamLink` is $1.021e{+}6$ (found at the end of the report file), which is about three times lower than the LR computed in `Familias`. This LR is calculated for all markers and replaces the LR calculated in `Familias`. Not to be confused with the multiplication factor that can be multiplied with the LR obtained in some other software.

f) There are four markers residing on chromosome 2, as well as three on chromosomes 5, 11 and 21.

Solution Exercise 4.11.

a) * Omitting details we get,

$$
\begin{aligned}
LR &= \frac{\Pr(\text{data} \mid H_1)}{\Pr(\text{data} \mid H_2)} \\
&= \frac{\begin{array}{l} 0.5 \cdot 4p_{12}p_{12}^{'}\Big(\big((1-r)^2 + r^2\big)4p_{21}p_{21}^{'} + 2r(1-r)p_{21}\Big) \\ + 0.5 \cdot p_{12}\Big(2r(1-r)4p_{21}p_{21}^{'} + ((1-r)^2 + r^2)p_{21}\Big)\end{array}}{4p_{12}p_{12}^{'} \cdot 4p_{21}p_{21}^{'}}
\end{aligned}
$$

where $p_{12}^{'}$ and $p_{21}^{'}$ is sloppy notation to indicate that this is the second time we observe this allele.

b) * The LR equals 2.058

c) ** Fortunately we do not need to concern ourselves with the alleles not shared between the individuals and therefore, $p_{12}^{'} = \theta + (1-\theta)p_{12}$ and $p_{21}^{'} = \theta + (1-\theta)p_{21}$ are the only two (updated) frequencies we have to compute. We may compute the updated frequencies also for the other alleles, but these will cancel out in the LR.

d) ** LR equals 1.72

e) ** See results in Figure 4.3.

4.2 X-chromosomal markers and FamLinkX

Solution Exercise 4.12.

a) The inheritance patterns are different for male and female meioses. While males pass on their only X-chromosome unchanged, the two X-chromosomes for females may recombine.

b) The recombination rate is 0.001 as can be found using `Tools -> cM ...`.

c) In order to account for linkage disequilibrium (association of alleles) we need to specify haplotype observations.

d) We find

$$r^2 = \frac{(p_{12}p_{16} - p_{12,16})^2}{p_{12}p_{13}p_{16}p_{17}} = \frac{(0.6 \cdot 0.6 - 59/100)^2}{0.6 \cdot 0.4 \cdot 0.6 \cdot 0.4} = 0.918.$$

which indicates a strong LD between the alleles.

g) LR (Exact)=99.97. Deviation from the theoretical value 100 is a consequence of the fact that λ is not exactly zero. `FamLinkX` does not allow the λ to be zero.

λ	LR
0.0001	13.01
0.01	12.98
1	10.69
100	1.50
10000	1.03

Table 4.1: Table of LRs for a number of different λ-s .

h) The LR changes dramatically. It now becomes 0.02 (LR(Exact)=0.21721, LR(cluster)=0.21709). The explanation is that given H_1 and disregarding mutations, the haplotypes for the child are fixed, while given H_2 other more common haplotypes are more probable. The consequence is that the likelihood is much lower given H_1 as this requires rare haplotypes for the child. Using a low value on λ we put almost all weight on the observations. Thus the haplotype observations will be crucial for the calculation of LR.

i) The answers may change quite a bit depending also on the choice of λ.

j) The degree of LD is extremely high which is evident from the results. It is also more probable that individuals actually share the most common haplotypes.

Solution Exercise 4.13.

a) The inheritance patterns differ. Two paternal female half siblings are obliged to share one allele IBD, whereas for maternal half siblings they may share one allele IBD with probability 0.5 and zero alleles IBD with probability 0.5.

d,e) Scaling versus Unrelated we find LR(Full siblings)=5050, LR(Maternal half siblings)=50.5 and LR(Paternal half siblings)=100. Looking at the LRs assuming LE, we see that the information in the haplotypes, and the underestimation of the LR, is great.

Solution Exercise 4.14.

c) The LRs are given in Table 4.1.

d,e) The LR approaches 13 as λ goes to 0 and 1 as λ goes to infinity, which is the LR when we do not account for haplotype observations. As λ

λ	LR
0.0001	0.8349
0.01	0.8348
1	0.8322
100	0.955
10000	1.270

Table 4.2: Table of LRs (Exact) for a number of different λ-s when the child is 13/13 for L1.

becomes big the expected haplotype frequencies are given much weight and dominate the haplotype probability estimates.

f) The LRs are given in Table 4.2. It seems that the value of λ does not influence the results considerably. Briefly, the explanation is that for H_1 we will sum over possible haplotypes for the founders and haplotypes with few observations and with many observations will be necessary to explain the data.

Solution Exercise 4.15.

b,c) The most probable relationships are given by

H_1 : The three females are all full siblings

H_2 : Two females are full siblings and the third (named F3) is a paternal half sibling

H_3 : Two females are full siblings and the third (named F3) is a maternal half sibling

When generating pedigrees in `Familias` we need some constraints. Otherwise the software will generate too many irrelevant pedigrees. Specifying all the typed females as children will create no pedigrees where they are parents to each other or other persons. Specifying the untyped persons as born the same year will create no pedigree where they are parent of each other.

g) Scaling versus H_2 we get an LR in favor of H_1 as $1.8e+11$ and an LR in favor of H_3 as $2.42e+5$. The LR comparing H_1 and H_3 is $7.44e+5$ in favor of the former hypothesis.

λ	LR
0.0001	4.9e-7
1	0.006
100	11.34
1000	217.93

Table 4.3: Table of LRs (exact)for a number of different values on λ

h) The final conclusion is that the data provide strong evidence in favor of the three females being full siblings, also compared to the next most probable hypotheses, i.e., H_3.

Solution Exercise 4.16.

c) The LRs are given in Table 4.3

d) We can conclude that for the range of λ–s considered, we obtain LRs that range from evidence against relationship to results that provide weak evidence in favor of relationship.

e) The answer can be found be exploring the frequency estimation tool. (Hint, found in the `Edit cluster` dialog.) We must further explore the hypotheses and see what haplotypes are necessary to explain the data. Given H_1 we see that the females share a common haplotype in each cluster, i.e., a certain haplotype can be distinguished. These haplotypes are rare, without any prior observations in the database. Given low values of λ, little weight will be given to unobserved haplotypes and they will have low frequencies. As a consequence the likelihood $\Pr(\text{data} \mid H_1)$ will be small, while the likelihood under H_2 will be higher as other, more common, haplotypes are more likely. In other words, without knowledge about the phase of what haplotypes are true under H_2, we must sum over all possible haplotypes.

f) It is indeed difficult to give a conclusion in the current case and to decide which λ to report. One may say the the evidence is inconclusive. We should further investigate if we are using an appropriate database, as the shared haplotype may be common in other populations. Fortunately, Example 7.4 explains how λ may be estimated.

Solution Exercise 4.17.

c) $LR(\text{Exact}) = 5.755e + 8$

d) $LR(\text{Exact}) = 7.310e + 6$

e) Given that λ equals to 1, LE model we will underestimate the evidence with a factor of 577/8.89=65. If we, on the other hand, use a λ of 212 (in this case, the size of the database), we get LRs that are close to each other, i.e., the difference between the model accounting for LD and the model assuming LE is small.

f) ** See Exercise 4.19

Solution Exercise 4.18.

d) The $LR(\text{Cluster}) = 4.5e - 5$ and the $LR(LE) = 1.95$

e) Tuning the value on λ, we see that when the value increases, $LR(\text{Cluster})$ approaches $LR(LE)$.

Solution Exercise 4.19.

a) The LR is computed as

$$LR = \frac{Pr(1/1) \cdot p_2}{Pr(1/1) \cdot Pr(1/2)} = \frac{1}{2 \cdot 0.4} = 1.25.$$

b) The LR is computed as

$$LR = \frac{p_1 \cdot p_2}{p_1 \cdot Pr(1/2)} = \frac{1}{2 \cdot 0.4} = 1.25.$$

d) $LR = 1.25 \times \frac{1}{0.6} = 2.08333$ in favor of paternity.

f) The LRs are given in Table 4.4

g) *The theoretical formula is derived below

$$LR = \frac{H_{1,3}H_{2,3}}{H_{1,3}2H_{2,3}H_{1,3}} = \frac{1}{2H_{1,3}}$$

λ	LR
0.01	49.89
1	40.73
100	4.00

Table 4.4: Table of exact LRs for different lambda-s.

where $H_{1,3}$ is the frequency of the haplotype with alleles 1 at L1 and 3 at L2. Using the the formula for haplotype frequency estimation we get that

$$LR = \frac{1}{2H_{1,3}} = \frac{1}{2 \cdot 0.125} = 4.$$

h) We use that

$$LR = \frac{1}{2H(1,3)}$$

and (by using the formula for haplotype estimation)

$$LR = \frac{C + \lambda}{2(c_i + p_i\lambda)}$$

where $C = 100$, $c_i = 1$ and $p_i = 0.4 \cdot 0.6 = 0.24$. Plotting functions are conveniently done in the open source software R, presented in detail in Chapter 5:

```
Function <- function(x) (100+x)/(2*(1+0.24*x))
curve(Function, 0, 1000, xlab="lambda",ylab="LR")
title(main="LR as a function of lambda")
grid()
Function(0.01)
Function(1)
Function(100)
```

Other software may also be used to produce the same plot. Figure 4.4 illustrates LR as a function of different values on λ.

i) ** The theoretical formula is derived below

$$LR = \frac{H_{12,17}H_{12,17}}{H_{12,17}H_{12,17}^2} = \frac{1}{H_{12,17}} = \frac{1}{0.01} = 100.$$

Solution Exercise 4.20.

a) The founder alleles are given by the alleles for the first father, the common mother and the second father. In total, there are 4 different alleles. The alleles for the fathers are given by the alleles of the sisters while the alleles for the mother are given by the alleles of the sisters as well as other possible alleles. In total, we have 16 different founder alleles sets. These are given by the sets [12 13 15 12], [12 15 13 12], [12 13 12 15], [12 12 13 15], [13 12 x 15], [13 x 12 15], [13 12 15 12] and [13 15 12 12], where x represents any of the five possible alleles.

b) There are now a number of possible founder alleles sets, for the fathers we still consider only the observed alleles, while for the mother we must consider the possibility of a mutation. The same sets as in a) are still possible, while in addition several other sets where the mother have alleles not observed in the two individuals are possible. For instance, the set [12 14 15 12] is possible, where a one step mutation must have produced the genotype for F1.

   We can use that for the sets [12 x y 12], [12 x y 15], [13 x y 15] and [13 x y 12] all values on x and y are possible with the exceptions that in the first case both cannot be 12, 13, 15 or 16 and combinations 12/13, 15/16 are not possible either; in the second case both cannot be 14, 15 or 16 and combinations 14/15, 14/16, 15/16 are not possible; in the third case both cannot be 14, 15 or 16 and combinations 14/15, 14/16, 15/16 are not possible; in the last case both cannot be 12, 13, 14, 15 or 16 and combinations 12/13, 14/15, 14/16, 15/16 are not possible. There are in total 100 different possible sets, if we subtract the sets that are not possible we get in total 100-27=73 different founder allele sets.

c) There are two meioses to account for, the two from the common mother to the two sisters.

d) There are $16 \cdot 2 = 32$ different combinations to consider.

e) There are $16 \cdot 2 \cdot 2 = 64$ different combinations to consider for the second marker

f) There are $16 \cdot 2 \cdot 2 \cdot 16 = 1024$ different combinations to consider for the second marker

g) There are $16 \cdot 2 \cdot 2 \cdot 16 \cdot 16 = 16,384$ different combinations to consider for the third marker

h) The unrelated persons can be treated separately and we have that the number of possible founder allele states for the combination of the three markers are given by the total number of different haplotype setups. In other words, there are $2 \cdot 2^3 = 16$ different combinations to consider for the third marker given H_2. Linkage/recombination is not a topic for unrelated individuals. Consider H_1, linkage has a minor effect in the current case, but given that many meioses are introduced, the number of computations will grow considerably. The main contributor to the number of different computations given H_1 is the possible founder alleles sets we must consider.

Solution Exercise 4.21.

d) The LR becomes 0.075, i.e. the data given the paternal half sibling relation is 1/0.075=13 times more probable.

f) The LR becomes 2.28 and the data therefore indicate that the maternal half sibling relation is twice as probable.

g) The LR becomes 0.21.

i) The LR in e) now becomes 0.47 while the LR in g) becomes 0.09

j) As we scale against the pedigree where the genetic inconsistency can be detected, i.e., paternal half siblings, the LR can be high given an "inappropriate" mutation model is selected. It should be noted that by swapping the hypotheses in LR formula, we would get an equally low LR. However, it is important to keep in mind that given the current example, how we model mutations is crucial to the conclusion.

Solution Exercise 4.22.

c) The LR becomes zero for all computation models.

d) A possible mutation is present at the marker DXS10101. A hint is to view and compare the data for both individuals in the Add DNA data window.

f) $LR(\text{Exact}) = 6.9e + 007$, the other computation models still yield an LR of zero as these do no consider mutations.

g) $LR(\text{Exact}) = 6.3e + 007$,

h) When we lower the value on the parameter, more"extreme" scenarios are considered, like double mutations etc, which individually provide a low likelihood but together contribute to the overall sum.

i) Same results as in g).

j) Again, same results as in g).

k) LR(*Step=2*)=119,793; LR(*Step=1*)=119,436; LR(*Step=0*)=117,957. It seems that whether the *Step* parameter is 1 or 2 only has a minor effect on the results, while lowering it to zero has a, in comparison, larger effect. However, the change may be considered as small compared to the large LRs.

l) The effect of the *Step* parameter is smaller unless two mutations are needed to explain the data. In the current case setting the step parameter to zero indicates that the common father under H_1 cannot possess any alleles other than the ones observed in the data. This restriction prohibits him from having some alleles.

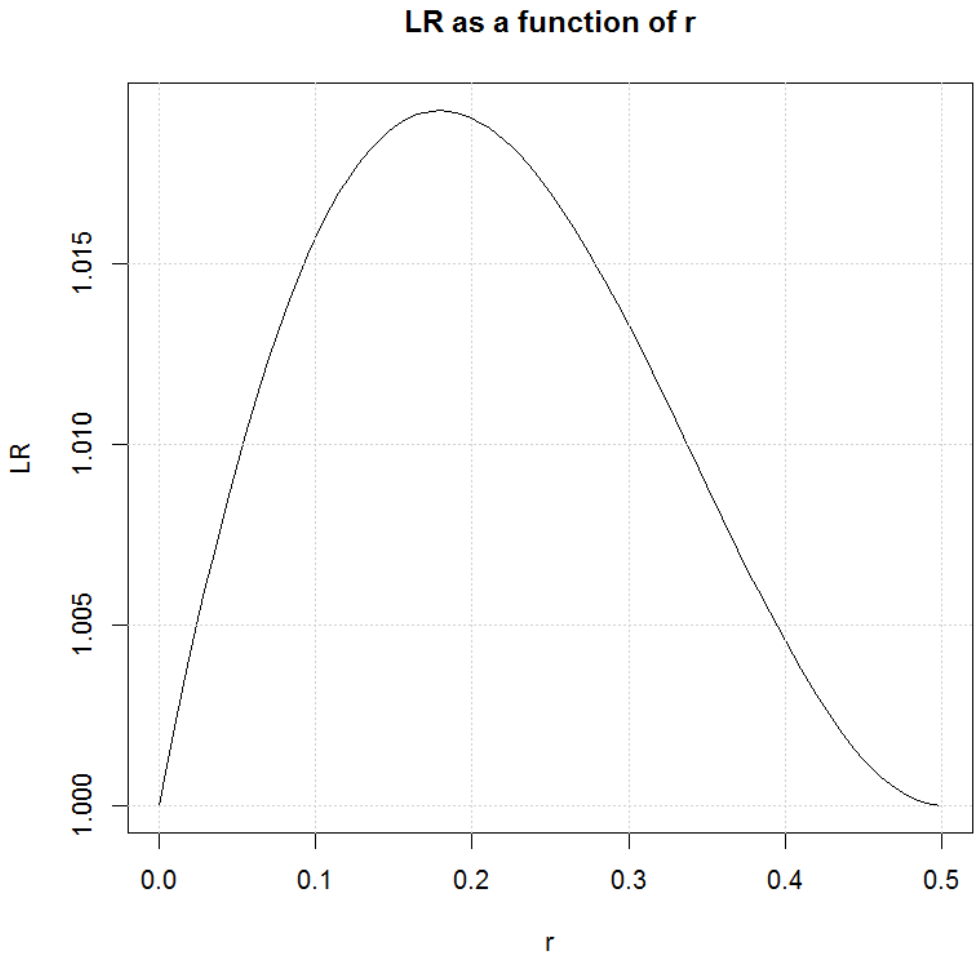


Figure 4.1: LR as a function of the recombination rate

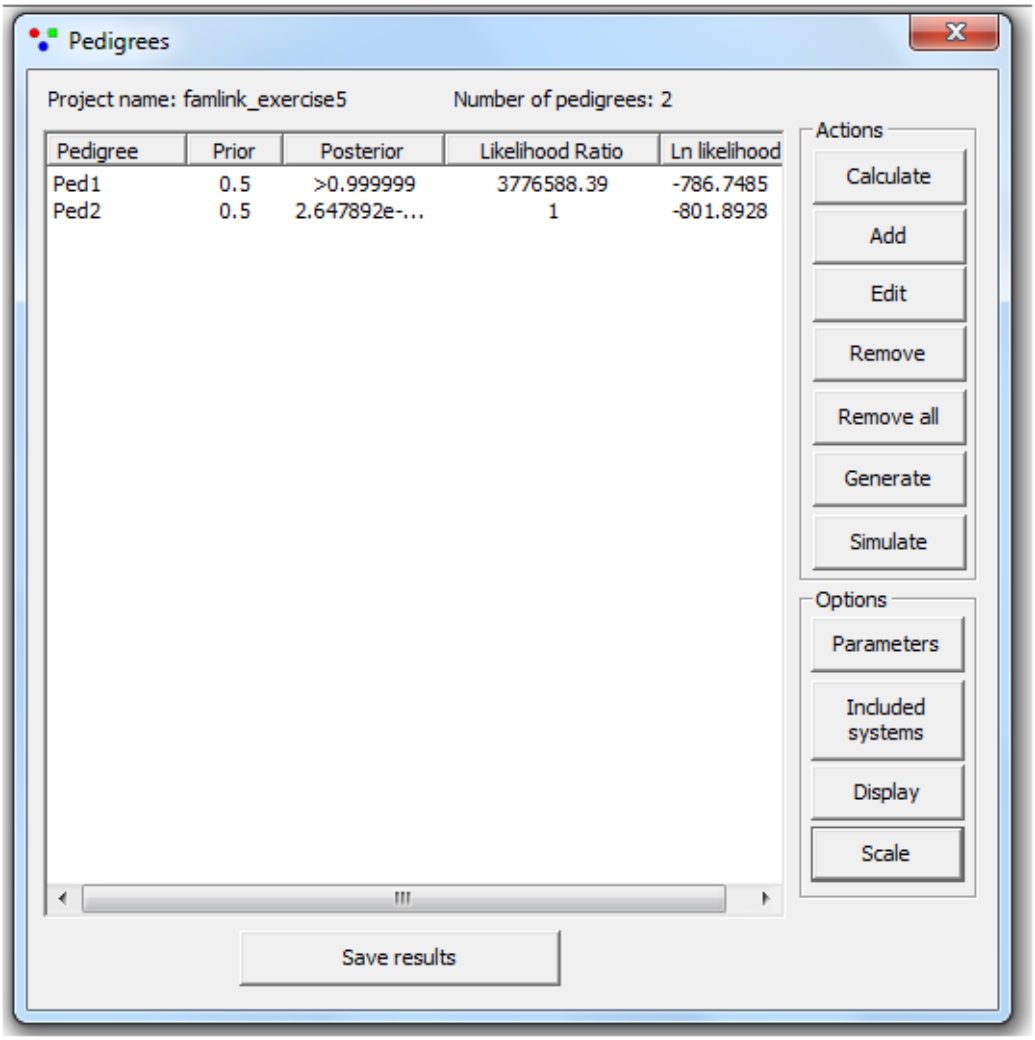


Figure 4.2: Results for Exercise 4.10

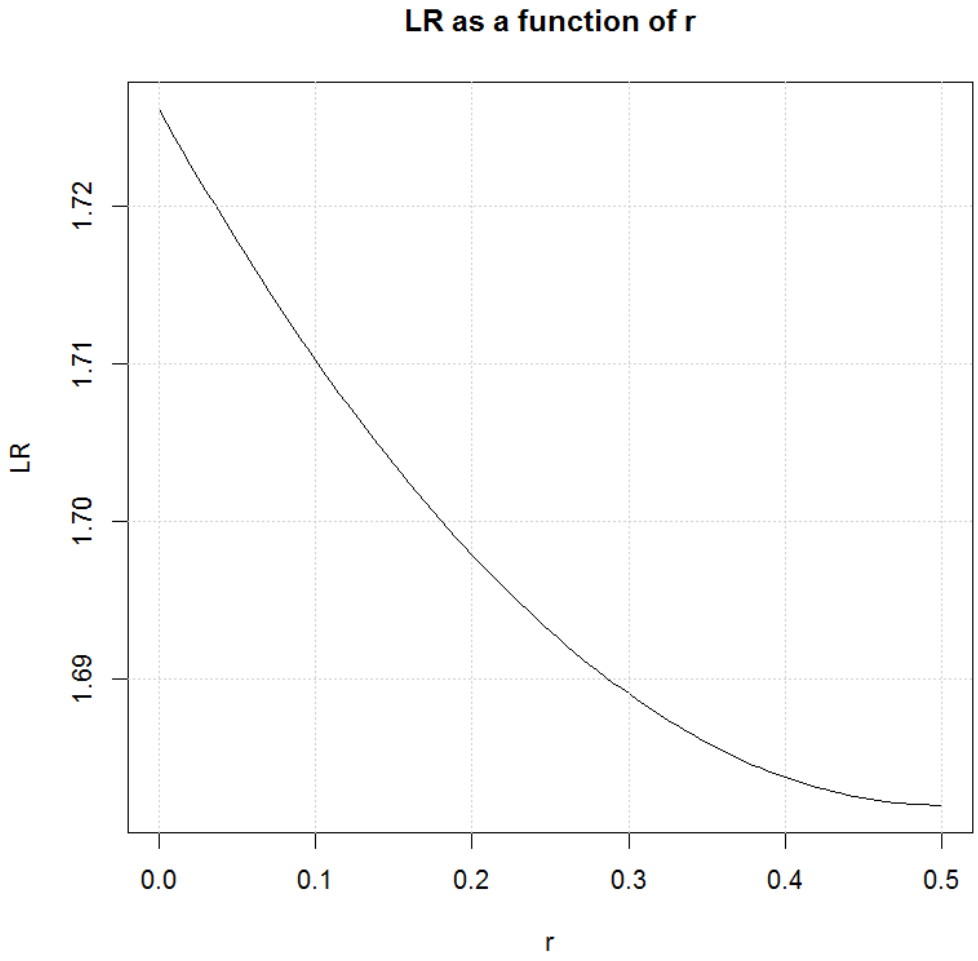


Figure 4.3: Results for Exercise 4.11 e)

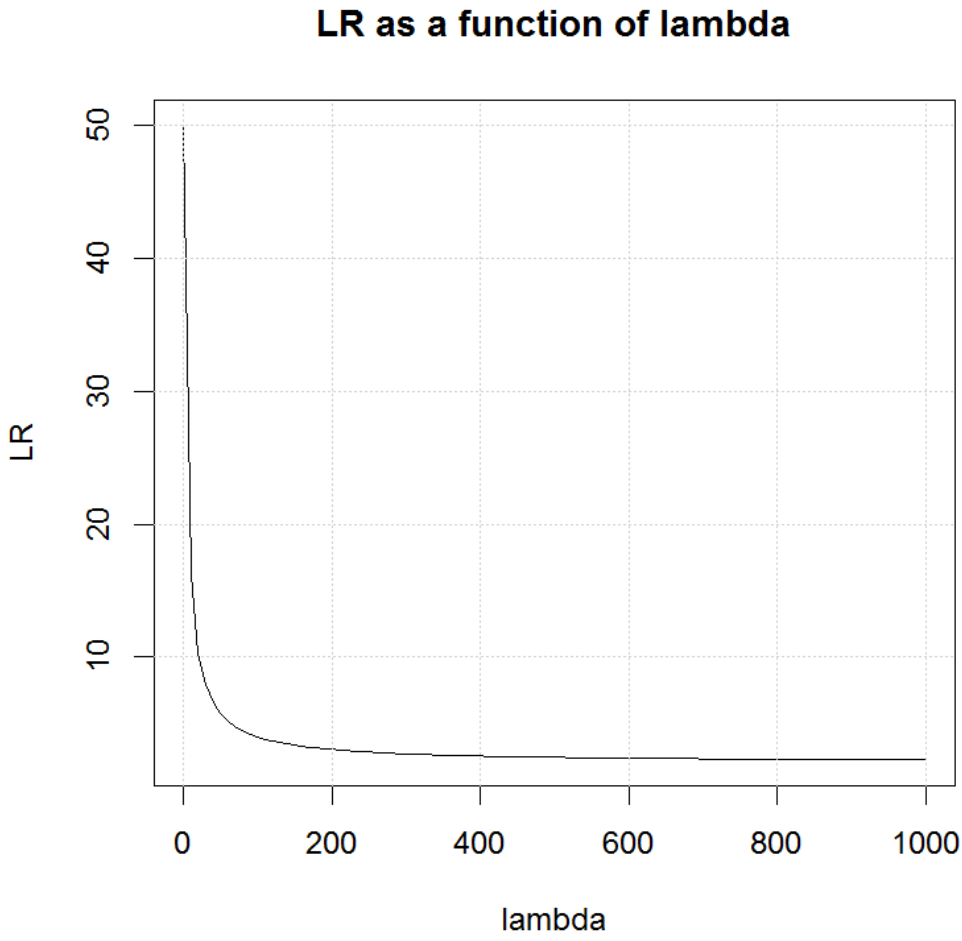


Figure 4.4: Plot of LR versus different values on λ.

Chapter 5

Solutions: "Relationship inference with R"

Solution Exercise 5.1.

a) We find

$$\begin{aligned} LR &= \frac{\Pr(\text{data} \mid \text{paternity})}{\Pr(\text{data} \mid \text{non-paternity})} \\ &= \frac{\Pr(\text{child} = A/C \mid \text{mother} = A/B, \text{AF} = C/D, \text{paternity})}{\Pr(\text{child} = A/C \mid \text{mother} = A/B)} \\ &= \frac{1/4}{0.3/2} = \frac{1}{0.6} = 1.666667. \end{aligned}$$

b) The object `result` contains a vector `posterior`, which lists the posterior probabilities of each of the pedigrees, in the same order as they were input, and assuming they have equal prior probabilities. We see that the posterior probability of the pedigree indicating paternity is 0.625. We also find the likelihood ratio for the second pedigree relative to the first equal to 1.666667, which is the same result as we obtained manually. The likelihoods 0.00144 and 0.00240 are also listed; these are the probabilities of observing all the genetic data specified, under the hypotheses of non-paternity or paternity, respectively.

c) For example

```
persons2 <- c("body", "B1", "B2", "father", "mother",
"grandma", "grandpa")
sex2 <- c("male", "male", "male", "male", "female",
 "female", "male")
ped3 <- FamiliasPedigree(id = persons2, dadid =
c("grandpa", "father", "father", "grandpa",
NA, NA, NA), momid = c("grandma", "mother", "mother",
"grandma", NA, NA, NA), sex = sex2)
ped4 <- FamiliasPedigree(id = persons2, dadid =
c(NA, "father", "father", "grandpa", NA, NA, NA),
momid = c(NA, "mother", "mother", "grandma",
NA, NA, NA), sex = sex2)
ped5 <- pedigree(id = persons2, dadid =
c("grandpa", "father", "father", "grandpa",
NA, NA, NA), momid = c("grandma", "mother",
"mother", "grandma", NA, NA, NA), sex = sex2)
ped6 <- pedigree(id = persons2, dadid = c(NA,
"father", "father", "grandpa", NA, NA, NA), momid =
c(NA, "mother", "mother", "grandma", NA, NA, NA),
sex = sex2)
plot(ped5)
plot(ped6)
```

d)
```
body <- c("A", "A")
B1   <- c("A", "B")
B2   <- c("A", "C")
datamatrix2 <- rbind(body, B1, B2)
pedigrees2 <- list(notuncle = ped4, uncle = ped3)
result2 <- FamiliasPosterior(pedigrees2, marker,
datamatrix2)
```

We get an LR in favour of the uncle hypothesis of 3.895833.

e) If the father does not have an A allele, his genotype must be B/C. Thus his possible genotypes are A/A, A/B, A/C, A/D, and B/C.

```
possibleGenotypes <- matrix(c(
"A", "A",
"A", "B",
```

```
"A", "C",
"A", "D",
"B", "C"), 5, 2, byrow = TRUE)
resultTable <- matrix(0, 5, 2)
for (i in 1:5) {
father <- possibleGenotypes[i,]
datamatrix3 <- rbind(datamatrix2, father)
resultTable[i,] <- FamiliasPosterior(pedigrees2, marker,
       datamatrix3)$likelihoods
}
```

We get

```
> resultTable
         [,1]       [,2]
[1,] 3.00e-06 9.0750e-05
[2,] 4.50e-06 1.2375e-05
[3,] 6.00e-06 1.6500e-05
[4,] 6.00e-06 1.6500e-05
[5,] 1.65e-05 4.1250e-06
> apply(resultTable, 2, sum)
[1] 0.00003600 0.00014025
```

These are the same results as should be obtained in the previous exercise.

f) First note that

$$\begin{aligned} & \Pr(\text{body} = A/A, \text{father} = A/A, \text{B1} = A/B, \text{B2} = A/C \mid \text{is uncle}) \\ = \; & \Pr(\text{B1} = A/B, \text{B2} = A/C \mid \text{father} = A/A) \\ & \cdot \Pr(\text{body} = A/A, \text{father} = A/A \mid \text{full brothers}) \end{aligned}$$

We see that in this case, the mother must have genotype B/C, so

$$\Pr(\text{B1} = A/B, \text{B2} = A/C \mid \text{father} = A/A) = 2 \cdot 0.2 \cdot 0.3 \cdot \frac{1}{2} \cdot \frac{1}{2} = 0.03$$

How to compute the last factor? One way is to use the IBD terminology:

$$\begin{aligned}
& \Pr(\text{body} = A/A, \text{father} = A/A \mid \text{full brothers}) \\
= \; & \Pr(\text{body} = A/A, \text{father} = A/A \mid \text{zero alleles IBD}) \\
& \cdot \Pr(\text{zero alleles IBD} \mid \text{full brothers}) \\
& + \Pr(\text{body} = A/A, \text{father} = A/A \mid \text{one allele IBD}) \\
& \cdot \Pr(\text{one allele IBD} \mid \text{full brothers}) \\
& + \Pr(\text{body} = A/A, \text{father} = A/A \mid \text{two alleles IBD}) \\
& \cdot \Pr(\text{two alleles IBD} \mid \text{full brothers}) \\
= \; & 0.1^4 \cdot \frac{1}{4} + 0.1^3 \cdot \frac{1}{2} + 0.1^2 \cdot \frac{1}{4} = 0.003025
\end{aligned}$$

The total result becomes $0.03 \cdot 0.003025 = 0.00009075$ which equals the result obtained above.

Solution Exercise 5.2.

a) The code for the other mutation models follows:

```
require(Familias)
R <- 0.005
r <- 0.5
persons <- c("CH", "AF")
sex <- c("male", "male")
ped1 <- FamiliasPedigree(id = persons, dadid <- c("AF", NA),
                         momid = c(NA, NA), sex = sex)
ped2 <- FamiliasPedigree(id = persons, dadid = c(NA, NA),
                         momid = c(NA, NA), sex = sex)
mypedigrees <- list(unrelated = ped2, isFather = ped1)
alleles <- 14:21
p <- c(0.072, 0.082, 0.212, 0.292, 0.222, 0.097, 0.020, 0.003)
CH <- c(16, 17)
AF <- c(14, 15)
datamatrix <- rbind(CH, AF)
mutmodels <- c("Equal", "Proportional", "Stepwise")
LR <- rep(NA, 3)
names(LR) <- mutmodels
theta <- 0
```

```
for(i in mutmodels[1:3]){
  locus1 <- FamiliasLocus(p, alleles, "locus1", MutationModel = i,
            MutationRange = r, MutationRate = R)
   LR[i] <- FamiliasPosterior(mypedigrees, locus1 ,
            datamatrix, kinship = theta)$LRperMarker[ , 2]
}
```

b) This follows from

```
pc <- 0.212
pd <- 0.292
m <- R/7
(1/2)*((pd+pc)*m)/(pc*pd)
```

c) Let (a, b, c, d) = (14, 15, 16, 17) and

```
mac <- 0.0012598425
mbc <- 0.0016842105
mad <- 0.0006299213
mbd <- 0.0008421053
```

Then the answer follows from

```
(1/4)*((mac + mbc)*pd+(mad+mbd)*pc)/(pc*pd)
```

d) Only one step mutations are allowed for this model. The required code follows:

```
M <- matrix(c(
0.9950,0.0050,0     ,0     ,0     ,0     ,0     ,0      ,
0.0025,0.9950,0.0025,0     ,0     ,0     ,0     ,0      ,
0     ,0.0025,0.9950,0.0025,0     ,0     ,0     ,0      ,
0     ,     0,0.0025,0.9950,0.0025,0     ,0     ,0      ,
0     ,     0,0     ,0.0025,0.9950,0.0025,0     ,0      ,
0     ,0     ,0     ,0     ,0.0025,0.9950,0.0025,0      ,
0     ,0     ,0     ,0     ,0     ,0.0025,0.9950,0.0025 ,
0     ,0     ,0     ,0     ,0     ,0     ,0.0050,0.9950),
 8, 8, byrow=TRUE)
```

```
locus1 <- FamiliasLocus(p, alleles, "locus1",
          MutationModel = "custom",
          MutationMatrix = M)
FamiliasPosterior(mypedigrees,
          locus1, datamatrix,
          kinship = theta)$LRperMarker[ , 2]
#  Using the formula we find the same answer:
mac <- 0
mbc <- 0.0025
mad <- 0
mbd <- 0
(1/4)*((mac + mbc)*pd+(mad+mbd)*pc)/(pc*pd)
```

	Equal	Proportional	Stepwise	Stationary	Custom
LR	0.00291	0.00626	0.00473	0.00640	0.00295

Table 5.1: LR for various mutation models in Exercise 5.2.

Solution Exercise 5.3. The solutions are essentially given in the exercise. There may some small differences in answers from the two versions as a result of rounding error.

Solution Exercise 5.4.

The code below does the job

a)
```
r <- seq(0, 0.5, length = 1000)
kappa1 <- (1 - r)/2
R <- r^2 + (1 -r)^2
kappa2 <- R/2
kappa3 <-((1-r)*R+r/2)/2
plot(r, kappa1, type = "l", lty = 1,
      ylab = "Pr(IBD = 1 for both markers)",
      xlab = "recombination rate")
lines(r, kappa2 , lty = 2)
lines(r, kappa3 , lty = 3)
legend (0.22, 0.5 , c("gf-gc","half-sibs","uncle-nep."),
         lty = 1:3)
```

b) Below, one large pedigree is plotted. Alternatively, three smaller plots can be made.

```
require(paramlink)
ped <- nuclearPed(2, sex = c(1, 2))
ped <- addOffspring(ped, mother = 4, noffs = 1)
ped <- addOffspring(ped, father = 5, noffs = 1)
plot(ped, title = "")
```

c) See derivation of (6.23) in the book. Possible code follpws:

```
p <- c(0.5, 0.5) #other values give same result
rho <- 0.29
g6 <- c(1, 1)
m1 <- marker(ped, alleles = 1:2, 6, g6, afreq = p)
m2 <- marker(ped, alleles = 1:2, 6, g6, afreq = p)
res <- twoMarkerDistribution(ped, 1, m1, m2, theta = rho)
numerator <- res["2/2", "2/2"]*p[1]^4
res <- twoMarkerDistribution(ped, 8, m1, m2, theta = rho)
denominator <- res["2/2", "2/2"]*p[1]^4
LR <- numerator/denominator
LR.formula <- (1-rho)/(rho^2+(1-rho)^2)
LR == LR.formula #TRUE
```

Solution Exercise 5.5.

A suggestion for solution is essentially provided in the exercise.

Solution Exercise 5.6.

A suggestion for solution is essentially provided in the exercise.

Solution Exercise 5.7.

a) The code follows:

```
require(disclapmix)
data(danes)
N <- sum(danes$n)
```

b) `p.count <- 1/(N + 1)`

c)
```
s <- sum(danes$n == 1L)
kappa <- (s + 1)/(N + 1)
p.brenner <- (1 - kappa)/(N + 1)
```

d)
```
danes_cor <- danes
danes_cor$DYS389II <- with(danes_cor, DYS389II - DYS389I)
danes_db <- as.matrix(danes_cor[rep(1L:nrow(danes_cor), danes_cor$n),
                        1 : 10 ])
dim(danes_db)[1]
```

e)
```
fit <- disclapmix(x = danes_db, clusters = 4L,
                  iterations = 500L)
singletons <- as.matrix(subset(danes_cor,
                         n == 1L)[, 1L : 10L ])
p.disclap <- predict(fit, newdata = singletons)
range(p.disclap)
```

f)
```
index.singletons <- (1:dim(danes_cor)[1])[ danes_cor$n == 1L ]
fit <- disclapmix(x = danes_db[ -index.singletons[1], ],
                  clusters = 4L, iterations = 500L)
predict(fit, newdata = matrix(
         danes_db[ index.singletons[1], ], nrow = 1 ))
```

Solution Exercise 5.8.

The code follows:

a)
```
require(DNAtools)
data(dbExample)
head(dbExample, 5)  # prints first 5 lines
tail(dbExample, 5)  # prints last 5 lines
str(dbExample)      # displays the structure of the database
phat <- freqEst(dbExample)
```

b)
```
barplot(phat[[ 1 ]], sub = names(phat.new)[[1]])
par(ask = TRUE)
lapply(phat, barplot)
par(ask = FALSE)
# Comment: Note that default ordering is not numerical and
# therefore allele 10 is sorted before 6 for TH01 as is seen from:
barplot(phat[[ 9 ]], sub = names(phat.new)[[ 9 ]])
```

c)
```
set.seed(123)
imdb <- dbSimulate(phat, theta = 0, n = 1000)
phat.new <- freqEst(imdb)
par(ask = TRUE)
lapply(phat, barplot)
par(ask = FALSE)
# Note that allele designations differ in the simulated database
# being 1,2, ...
```

Chapter 6

Solutions: "Models for pedigree inference"

Solution Exercise 6.1 (Properties of the mutation matrix)**.**

Note that

$$M_f \mathbf{1}^t = (1-c) I \mathbf{1}^t + c \mathbf{1}^t p \mathbf{1}^t = (1-c)\mathbf{1}^t + c\mathbf{1}^t = \mathbf{1}^t$$

as required. The diagonal elements of M_f are $(1-c)+cp_i$ and the off digonal, cp_i, all positive. Stationarity follows from

$$pM_f = p(1-c)I + pc\mathbf{1}^t p = (1-c)p + cp = p.$$

By definition,

$$R = 1 - \sum_{i=1}^{n} m_{ii} p_i = 1 - \sum_{i=1}^{n} \left((1-c)p_i + cp_i^2 \right) = 1 - \mathrm{Tr}(D(p)M).$$

Solution Exercise 6.2 (Mixtures and relatives)**.**

a) Enter

```
E <- 1:3
datamatrix <- generate( E, K = NULL, 2)
```

b) Revise the code below.

c,d) The complete code is

```
require(Familias)
require(BookEKM)
persons <- c("CH", "MO", "AF")
ped1 <- pedigree( id = persons, dadid = c( "AF", NA, NA),
                  momid = c( "MO", NA, NA),
sex <- c( "male", "female", "male"))
ped2 <- pedigree( id = c( persons, "TF"),
                  dadid = c( "TF", NA, NA, NA),
                  momid =c ( "MO", NA, NA, NA),
                  sex =c ( "male", "female", "male", "male"))
pedigrees <- list( isFather = ped1, unrelated = ped2)
E <- 1:3 ; gAF <- c( 3,4)
datamatrix <- generate( E, K = NULL, 2)
AF <- rep( gAF, dim(datamatrix)[2]/2)
datamatrix <- rbind(datamatrix,AF)
datamatrix <- as.data.frame(datamatrix)
rownames(datamatrix)[c(1,2)] = c( "CH", "MO")
R <- 0.00
locus <- FamiliasLocus( frequencies = rep( 0.2, 5),
                        allelenames = c( 1:5), name = "V1",
                        MutationRate = R)
theta <- seq(0, 0.1, length = 100)
LR <- NULL
for (i in theta)
   LR <- c(LR,mix3Familias( pedigrees, locus, datamatrix,kinship=i)$LR)
plot(theta, LR, type = "l")
```

Remaining solutions are modifications of the above ones.

e) We find $LR = 2.5$.

Solution Exercise 6.3 (Dropout. Models and interpretations)**.**

a) Using Bayes theorem we find

$$\begin{aligned}\Pr(D = 1 \mid \text{data} = \text{a}/-) &= \frac{p_a 2d(1-d)}{(1-d)^2 p_a^2 + 2p_a d(1-d)} \\ &= \frac{2d}{(1-d)p_a + 2(1-p_a)d} \to 1 \text{ as } \text{p}_\text{a} \to 0\end{aligned}$$

b) We find

$$\Pr(a/b \mid \text{data} = \text{a}/-) = \frac{d(1-d)2p_a(1-p_a)}{(1-d)^2p_a^2 + 2p_a(1-p_a)d(1-d)} = \frac{2d(1-p_a)}{(1-d)p_a + 2d(1-p_a)}.$$

Solution Exercise 6.4 (Inbreeding. Jacquard. `paramlink`)**.**

a) These states are 0 as the probability of IBD within an individual is 0.

b) The IBD coefficients for full sibs are 1/4, 1/2 and 1/4 corresponding for sharing 0, 1 or 2 alleles, from which the first part follows. The second part is a consequence of the definition of IBD.

c) The below code produces the required plots:

```
require(paramlink)
alleles <- c('a', 'b')
H2 <- nuclearPed(2)
m2 <- marker(H2, alleles=alleles, 3:4, c('a','a'))
H1 <- addParents(H2, 1, father=10, mother=11)
H1 <- addParents(H1, 2, father=10, mother=11)
m1 <- marker(H1, alleles=alleles, 3:4, c('a','a'))
plot(H1, marker=m1)
plot(H2, marker=m2)
```

d)
```
p1 <- c(0.1, 0.9)
m1.empty <- marker(H1, alleles = alleles, afreq = p1)
oneMarkerDistribution(H1, ids = c(3,4), partial = m1.empty,
                      loop_breaker = 1)
```

e) Some algebra shows this; a numerical check:

```
p <- p1[ 1]
Delta <- c(2, 1, 4, 1, 4, 1, 7, 10, 2)/32
g <-  c(p, p^2, p^2, p^3, p^2, p^3, p^2, p^3, p^4)
sum(Delta*g)
(1+8*p+6*p^2+p^3)*p/16
```

f) This follows from above previous expressions. The code for the plot:

```
p <- seq(0.001, 0.1, length = 1000)
LR <- (1/4)*(1+8*p+6*p^2+p^3)/(p+2*p^2+p^3)
plot(p, log(LR,base = 10), type = "l")
```

g) The code and a check follow:

```
m1 <- marker(H1, alleles = alleles, afreq = p1, 3, c("a","a"))
oneMarkerDistribution(H1, 4, m1, loop_breakers = 1)
# For instance Pr(a/a|a/a)
0.0116/(0.0116+0.0061+0.0147)
```

h) For instance, we may enter

```
Nsim <- 10000
res <- markerSim(H1, N = Nsim, available = 4,
                   partial = m1, loop_breaker = 1)
genotypes <- as.data.frame(res, singleCol = TRUE, sep= "/")
geno4 <- as.character(genotypes[6, -(1:5)])
table(geno4)/Nsim
```

Solution Exercise 6.5 (Jacquard coefficients. `identity`).

Possible code is given below:

a)
```
install.packages("identity")
require(identity)
ped <- rbind(c(1, 0, 0),
             c(2, 0, 0),
             c(3, 1, 2),
             c(4, 1, 2),
             c(5, 3, 4),
             c(6, 3, 4))
identity.coefs(c(5, 6), ped) #  Pedigree H1
identity.coefs(c(3, 4), ped) #  Pedigree H2
```

b)
```
ped <- rbind(c(1, 0, 0),
             c(2, 0, 0),
             c(3, 2, 1),
```

```
             c(4, 3, 1))
Delta <- identity.coefs(c(1, 4), ped)[ 2, -c(1, 2)]
p <- 0.1
g <- c(p, p^2, p^2, p^3, p^2, p^3, p^2, p^3, p^4)
sum(Delta*g)

require(Familias) #Checking with Familias)
persons <- c("I.1", "I.2", "II.2", "III.1")
sex <- c("male", "female", "female", "male")
ped1 <- pedigree(id = persons, dadid =c(NA, NA, "I.1", "I.1"),
              momid = c(NA, NA, "I.2", "II.2"), sex = sex)
locus1 = FamiliasLocus(c(p, 1-p), c("a", "b"))
I.1 <- III.1 <- c("a", "a")
datamatrix <- rbind(I.1, III.1)
FamiliasPosterior(ped1, locus1, datamatrix)

require(paramlink) # Checking with paramlink
x <- nuclearPed(1, sex = 2)
x <- addOffspring(x, mother = 3, father = 1, noffs = 1,sex = 1)
x <- breakLoops(x, loop_breakers = 3)
m <- marker(x, alleles =c("a","b"), 1, c("a", "a"),
            4, c("a", "a"), afreq = c(p, 1-p))
likelihood(x, m)
```

Solution Exercise 6.6.

A suggestion for solution is essentially provided in the exercise.

Chapter 7

Solutions: "Parameter Estimation and Uncertainty"

Solution Exercise 7.1 (7.1: Estimation of kinship)**.**

The code for the calculation in `Familias` could be (see plot for explanation of abbreviations):

```
require(Familias)
persons <- c("gf" , "gm" , "mo1","so1","so2", "mo2","co1","co2")
sex  <- c("male", "female","female", "male", "male", "female", "male", "male")
dadid <-   c(NA,  NA , NA, "gf" ,"gf",  NA, "so1", "so2")
momid <-   c(NA,  NA, NA, "gm" ,"gm",  NA, "mo1", "mo2")
cousins <- FamiliasPedigree(id = persons, dadid = dadid, momid = momid, sex = sex)
plot(cousins)
dadid <-   c(NA,  NA , NA, "gf" ,"gf",  NA, NA, NA)
momid <-   c(NA,  NA, NA, "gm" ,"gm",  NA, "mo1", "mo2")
unrelated <- FamiliasPedigree(id = persons, dadid = dadid, momid = momid, sex = sex)
pedigrees <- list(isCousins = cousins, unrelated = unrelated)
data(NorwegianFrequencies)
D21S2055 <- FamiliasLocus(NorwegianFrequencies$D21S2055,
                          name = "D21S2055", MutationRate = 0.005,
                          MutationModel = "Proportional")
co1 <- c(28,28)
co2 <- c(28,28)
datamatrix <- rbind(co1, co2)
kinship1 <- 0
```

```
kinship2 <- 0.1
LR1 <-FamiliasPosterior(pedigrees, D21S2055, datamatrix,ref=2,
                        kinship=kinship1)$LR["isCousins"]
LR2 <-FamiliasPosterior(pedigrees, D21S2055, datamatrix,ref=2,
                        kinship=kinship2)$LR["isCousins"]

#Simulation alternative
N <- 4000
res <- kinshipBySimulation(pedigrees, D21S2055, datamatrix, kinship1, N)
LR3 <- res$likelihoods[1]/res$likelihoods[2]
res <- kinshipBySimulation(pedigrees, D21S2055, datamatrix, kinship2, N)
LR4 <- res$likelihoods[1]/res$likelihoods[2]
c(LR1,LR3,LR2, LR4)
```

Chapter 8

Solutions: "Making Decisions"

Solution Exercise 8.1.

a) The optimal decision is

$$\begin{array}{ll} \text{choose } H_1 & \text{if } o > c_2 \\ \text{choose } H_2 & \text{if } o < \frac{1}{c_1} \\ \text{make no decision} & \text{otherwise} \end{array}$$

As $o = \text{LR} \cdot o_{0,\text{standard}}$, this decision rule can be rewritten as

$$\begin{array}{ll} \text{choose } H_1 & \text{if LR} > \frac{c_2}{o_{0,\text{standard}}} \\ \text{choose } H_2 & \text{if LR} < \frac{1}{c_1 \cdot o_{0,\text{standard}}} \\ \text{make no decision} & \text{otherwise} \end{array}$$

The lab's decision rule is

$$\begin{array}{ll} \text{choose } H_1 & \text{if LR} > C_H \\ \text{choose } H_2 & \text{if LR} < C_L \\ \text{make no decision} & \text{otherwise} \end{array}$$

If this is the optimal rule it follows that

$$\begin{aligned} C_H &= \frac{c_2}{o_{0,\text{standard}}} \\ C_L &= \frac{1}{c_1 \cdot o_{0,\text{standard}}} \end{aligned}$$

giving

$$\begin{aligned} c_1 &= \frac{1}{C_L \cdot o_{0,\text{standard}}} \\ c_2 &= C_H \cdot o_{0,\text{standard}} \end{aligned}$$

b) Replacing c_1 and c_2 in

$$\frac{1}{c_1} < o_0 \cdot x < c_2$$

with their computed values yields

$$C_L \cdot o_{0,\text{standard}} < o_0 \cdot x < C_H \cdot o_{0,\text{standard}}$$

or

$$\frac{o_{0,\text{standard}}}{o_0} C_L < x < \frac{o_{0,\text{standard}}}{o_0} C_H.$$

As $\frac{o_{0,\text{standard}}}{o_0} > 1$ it is clear there exists x which satisfy the inequalities above and also $x > C_H$.

c) The expected cost of deciding for H_1 is

$$\Pr(H_2) \cdot (1 + c_2) = \frac{1}{1+o}(c_2 + 1) = \frac{1 + o_{0,\text{standard}} \cdot C_H}{1 + o_0 \cdot \text{LR}}$$

As we assume that LR satisfies

$$o_0 \cdot \text{LR} < o_{0,\text{standard}} \cdot C_H$$

it follows that the cost is always larger than 1.

Solution Exercise 8.2.

a) The cost of deciding on H_1 when H_2 is true was denoted as $1 + c_2$ in the text. It was also shown that $c_2 = L_H$, when cutoff rates reflect optimal decisions. Thus, the expected cost of deciding on H_1 is

$$\Pr(H_2 \mid D)(1 + L_H)$$

The posterior probability for H_2 can be expressed in terms of the posterior odds o using

$$\Pr(H_2 \mid D) = \frac{1}{1+o}$$

and we also have that $o = \mathrm{LR} \cdot o_0 = \mathrm{LR}$ with the prior odds $o_0 = 1$. Putting this together, the expected cost of deciding on H_1 becomes

$$\Pr(H_2 \mid D)(1 + L_H) = \frac{1 + L_H}{1 + \mathrm{LR}}$$

Similarly, the expected cost of deciding on H_2 is

$$\Pr(H_1 \mid D)(1+c_1) = \frac{o}{1+o}(1+1/L_L) = \frac{\mathrm{LR}}{1+\mathrm{LR}}(1+1/L_L) = \frac{\mathrm{LR}+\mathrm{LR}/L_L}{1+\mathrm{LR}}.$$

b) The costs of the three possible decisions are

$$\begin{aligned} \text{Deciding on } H_1\text{:} \quad & \frac{1 + L_H}{1 + \mathrm{LR}} \\ \text{Deciding on } H_2\text{:} \quad & \frac{\mathrm{LR}+\mathrm{LR}/L_L}{1+\mathrm{LR}} \\ \text{Making no decision:} \quad & 1 \end{aligned}$$

In the case that $0 < \mathrm{LR} < L_L$ we get that

$$\frac{\mathrm{LR}+\mathrm{LR}/L_L}{1+\mathrm{LR}} < \frac{\mathrm{LR}+1}{1+\mathrm{LR}} = 1 < \frac{1+L_H}{1+\mathrm{LR}}$$

so choosing H_2 minimizes the cost. The extra costs associated with taking the other decisions can be found by computing the differences of the costs above and are given in Table 8.1.

In the case that $L_L \leq \mathrm{LR} \leq L_H$ we get that

$$1 \leq \frac{1+L_H}{1+\mathrm{LR}}$$

and

$$1 \leq \frac{\mathrm{LR}+\mathrm{LR}/L_L}{1+\mathrm{LR}}$$

so making no decisions is optimal. Extra costs of other decisions are given in Table 8.1.

Finally, when $L_H < \mathrm{LR}$, we get

$$\frac{1+L_H}{1+\mathrm{LR}} < \frac{1+\mathrm{LR}}{1+\mathrm{LR}} = 1 < \frac{\mathrm{LR}+\mathrm{LR}/L_L}{\mathrm{LR}+1}$$

so H_1 is the optimal decisions, and additional costs of other decisions are again found in Table 8.1.

	$0 < \text{LR} < L_L$	$L_L \leq \text{LR} \leq L_H$	$L_H < \text{LR}$
Standard rule	0	0	0
Exclusion rule (RMNE $< o_0/c_2$)	$\frac{1+L_H-\text{LR}-\text{LR}/L_L}{1+\text{LR}}$	$\frac{L_H-\text{LR}}{1+\text{LR}}$	0
Exclusion rule (RMNE $\geq o_0/c_2$)	$\frac{1-\text{LR}/L_L}{1+LR}$	0	$\frac{\text{LR}-L_H}{1+\text{LR}}$

Table 8.1: In each of the four cases given in the top row, the table shows the additional cost of following a decision rule compared to the cost of following the optimal rule.

Solution Exercise 8.3.

a) Assuming H_2, AF and CH are unrelated, and the result follows.

b) This is achieved by:

```
require(paramlink)
ped <- nuclearPed(1)
alleles <- c("a", "b")
afreq <- c(0.5, 0.5)
m <- marker(ped, alleles = alleles, afreq = afreq)
tableHD <- oneMarkerDistribution (ped, c(1, 2), m)
tableHP <- oneMarkerDistribution (ped, c(1, 3), m)
```

c) We find $\Pr(\mathbb{X}_1 = 0) = 0$ and $\Pr(\mathbb{X}_2 = 0) = 0.0625 + 0.0625 = 0.125$.

d) The previous answer gives one entry in the table, the other entries can be calculated similarly.

e) We find

$$E[\mathbb{X}_1^{-1}] = 0.75 + \frac{1}{2} \times 0.25 = 0.875 = \Pr(\mathbb{X}_1 > 0).$$

Solution Exercise 8.4.

a) We find

$$E(\mathbb{X}_1) = 0 \times 0.000 + 1 \times 0.75 + 2 \times 0.250 = 1.25.$$
$$E(\mathbb{X}_2) = 1 \times 0.125 + 1 \times 0.75 + 1 \times 0.125 = 1.00.$$

With $k_1 = 1$ and $k_2 = 0$

$$E[\mathbb{X}_1] = \frac{1}{4}L + (1 - 0 - \frac{1}{4}) = \frac{L+3}{4} = \frac{5}{4}$$

b) This follows from the expression for the expectation: $a_1L^2 + b_1L + c_1 > b_2L + c_2$ when $L > L_0$ for some L_0 when $a_1 > 0$.

Solution Exercise 8.5.

a) $PE = 0.0512$.

b) Required additional code:

```
p <- c(0.2, 0.8)
exclusionPower(claim, true, available, alleles = 2, afreq = p,
               known_genotypes <- list(c(3, 1, 1)))
```

c) Required additional code:

```
exclusionPower(claim, true, available, alleles = 2,
               afreq = p, Xchrom = TRUE)
```

d) Required additional code:

```
p <- c(0.7, 0.1, 0.1, 0.1)
exclusionPower(claim, true, available, alleles = 1:4, afreq = p)
exclusionPower(claim, true, available, alleles = 1:4, afreq = p,
               known_genotypes = list(c(3, 1, 1)))
exclusionPower(claim, true, available, alleles = 1:4,
               afreq = p, Xchrom = TRUE)
```

e) Required code:

```
mother.daughter <- nuclearPed(1, sex = 2)
sisters <- relabel(nuclearPed(2, sex = c(2, 2)), c(101, 102, 2, 3))
PE1 <- exclusionPower(ped_claim = mother.daughter,
                  ped_true = sisters, ids = c(2, 3), alleles = 2)
```

f) Required code for last part:

```
sisters.LOOP <- addParents(sisters, 101, father = 201, mother = 202)
sisters.LOOP <- addParents(sisters.LOOP, 102, father = 201, mother = 203)
exclusionPower(ped_claim = mother.daughter, ped_true = sisters.LOOP,
               loop = 101,ids = c(2, 3), alleles = 2,
               afreq=c(0.1, 0.9), known_genotypes=list(c(3, 1, 1)))
```

Solution Exercise 8.6.

a) Answer provided, i.e.,

```
x <- nuclearPed(2, sex=1)
data(NorwegianFrequencies)
L1 <- NorwegianFrequencies[["SE33"]]
m <- marker(x, alleles=names(L1), afreq=L1,
            3, c(11,12), 4, c(12,13))
simPed <- markerSim(x, N=5, available = c(1,2),
                    partialmarker = m, seed = 17,verbose=FALSE)
```

b) `plot(simPed, marker = 1:5)`

Solution Exercise 8.7.

a) Answer provided

b) See below.

c)

```
set.seed(1234)
nsim <- 10000
x <- sample.profiles(N = nsim, freqs = freqsNLngm)
x.FS <- sample.relatives(x, 1, type = "FS")
SI <- ki(x, x.FS, hyp.1="FS", hyp.2="UN", freqs = freqsNLngm)
range(SI)  #Comment: extreme variation
hist(log(SI, base = 10),
     xlab = "log10(SI)", prob = TRUE)
length(SI[ SI < 1 ])/nsim  # estimates P(LR < 1| sibs)
```

Solution Exercise 8.8.

a)

```
t <- 5
p <- 1 - pnorm(t)
```

b)

```
set.seed(17)
N <- 10^6
z <- rnorm(N, mean = 0)
p1.hat <- mean(z > t)
```

c)

```
set.seed(17)
z <- rnorm(N, mean = t)
p2.hat <- mean((z > t)*dnorm(z)/dnorm(z, mean = 5))
```

d) Answer provided.

e)

```
require(DNAprofiles)
data(freqsNLngm)
hp <- ki.dist(hyp.1 = "PO", hyp.2 = "UN", hyp.true = "PO",
              freqs.ki = freqsNLngm)
hd = ki.dist( hyp.1 = "PO", hyp.2 = "UN", hyp.true = "UN",
              freqs.ki = freqsNLngm)
set.seed(100)
q <- sim.q(t = 0, dists = hd, dists.sample = hp, N = 1e5)
# Exact value
prod(1-sapply(hd, function(y) y$fx[ 1])) #Exact
#Alternative:
pair.H1 <- dists.product.pair(hd, appr = TRUE)
cdf.H1 <- dist.pair.cdf(pair.H1)
1 - cdf.H1(0)
```

f)

```
hp <- ki.dist(hyp.1 = "PO", hyp.2 = "UN", hyp.true = "PO",
              freqs.ki = freqsNLngm[ 1])
hd <- ki.dist(hyp.1 = "PO", hyp.2 = "UN", hyp.true = "UN",
              freqs.ki = freqsNLngm[ 1])
LR <- hd[[ 1]]$x
fx <- hd[[ 1]]$fx
t <- quantile(LR, probs = 0.5)
```

```
q.exact <- sum(fx[ LR > t ])
nsim <- 100
q <- rep(NA, nsim)
for (i in 1:nsim)
 q[ i] <- sim.q(t, dists = hd, dists.sample = hp, N = 1e5)

plot(density(q), xlab = "Estimated q",
      main = "Density estimate of exceedance probability
      \n exact: vertical line")
abline(v = q.exact)
```

Solution Exercise 8.9.

a) The four probabilities of the example are estimated as follows

```
require(Familias); require(DNAprofiles)
data(NorwegianFrequencies)
set.seed(17);N=1e6
h1 <- ki.dist(hyp.1 = "FS", hyp.2 = "UN", hyp.true = "FS",
              freqs.ki = NorwegianFrequencies[1:15])
h2 <- ki.dist(hyp.1 = "FS", hyp.2 = "UN", hyp.true = "UN",
              freqs.ki = NorwegianFrequencies[1:15])

p1 <- 1-sim.q(t=10000, N=N, dists=h1)
p2 <- 1-sim.q(t=1/10000, N=N, dists=h1)
p3 <- sim.q(t=1/10000, N=N, dists=h2)
set.seed(17)
p4 <- sim.q(t=10000, N=N, dists=h2, dists.sample = h1)
c(p1, p2, p3, p4)
```

b) Expand on the previous code as below:

```
set.seed(17)
p.with <- p.with.out <- NULL
for (i in 1:10){
p.with <- c(p.with,sim.q(t=10000, N=N, dists=h2, dists.sample = h1))
p.with.out <- c(p.with.out,sim.q(t=10000, N=N, dists=h2))
}
res <- cbind(p.with,p.with.out)
apply(res,2,mean)
```

The average values based on the simulations agree well, but there is less variability, as expected, when importance sampling is used.

c) Code follows:

```
expectedCostExample.8 <-function( L.L = 1/10000,L.H = 10000, odds,
                          c.D , p1, p2, p3,p4 )
    c.D + odds/(odds + 1) * (p1 + p2/L.L) + 1/(odds + 1) *
   (p3 + p4 * L.H)
expectedCostExample.8(c.D=0.5,odds=100,p1=0.65,p2=0,p3=1,p4=9.83*10^{-4})
```