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

Solutions X Exercises 4.12-4.22

Thore Egeland
Daniel Kling
Petter Mostad

August 24, 2016

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

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

c) Answer as below as linkage does not matter for this case.

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 additition 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.
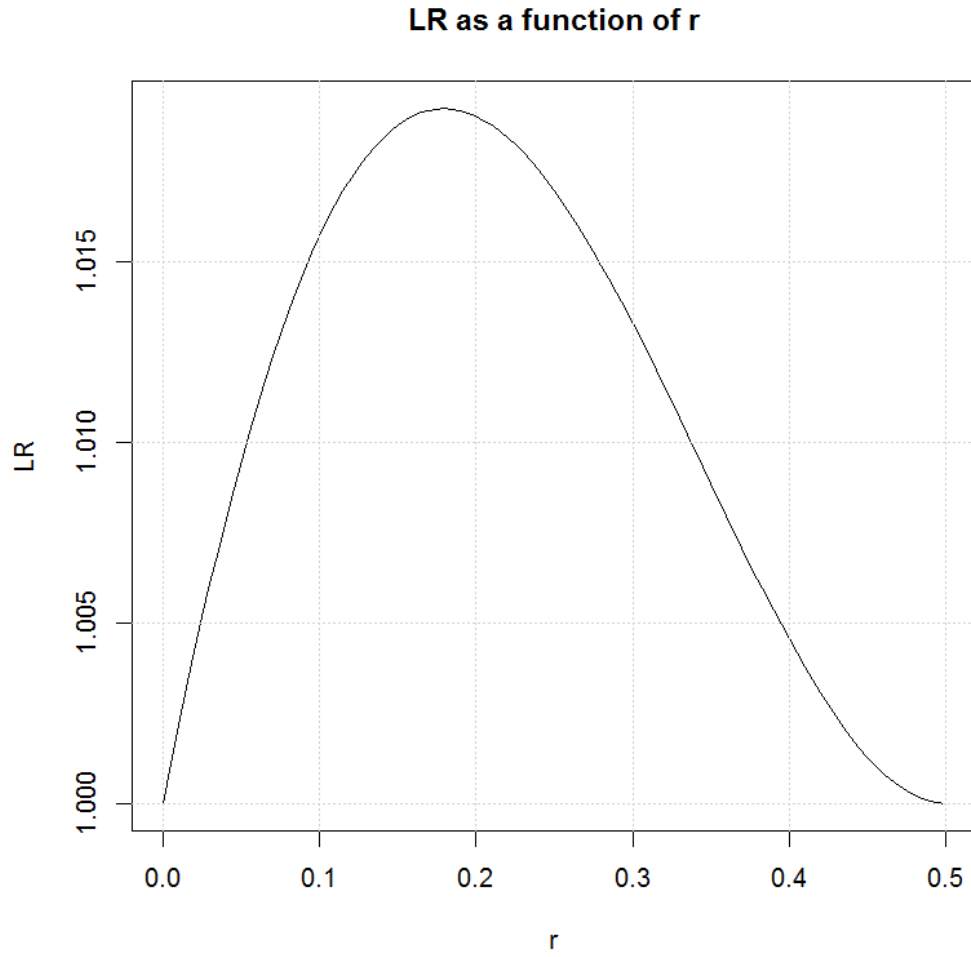
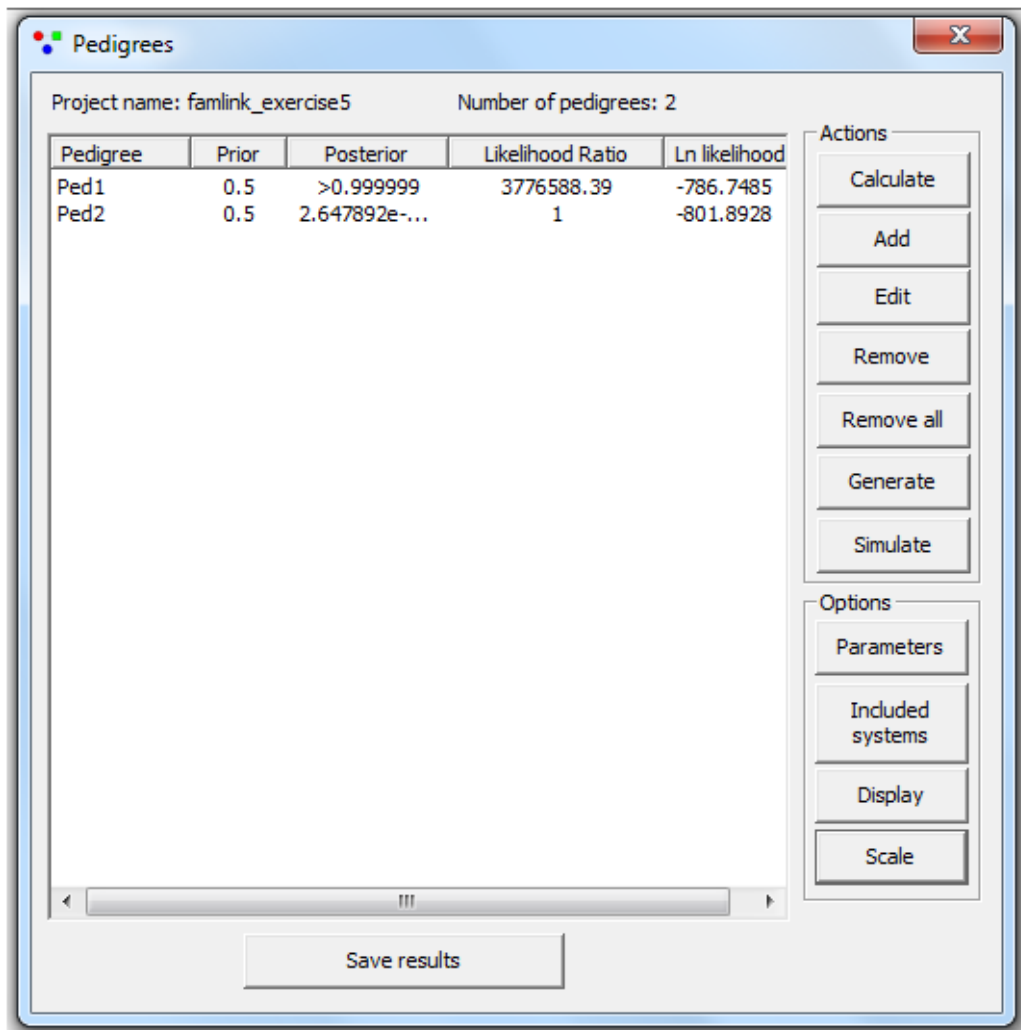Figure 4.1: LR as a function of the recombination rate
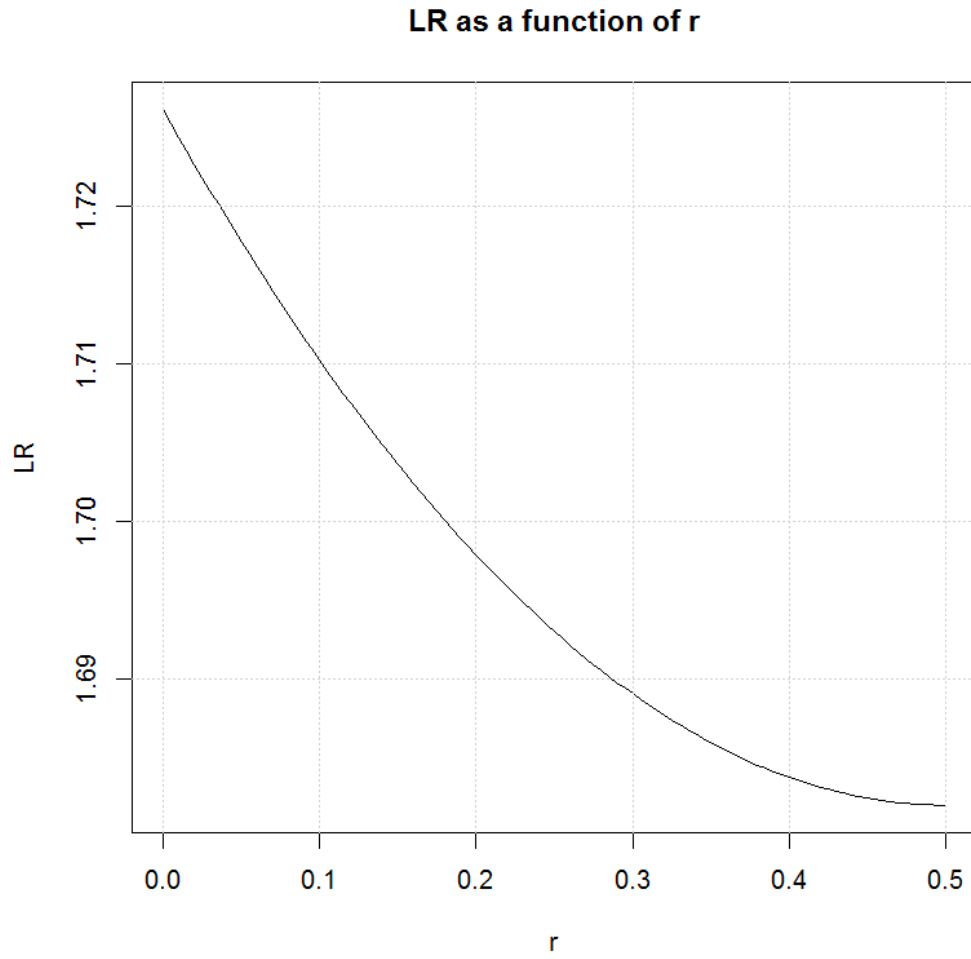
Figure 4.2: Results for Exercise 4.10

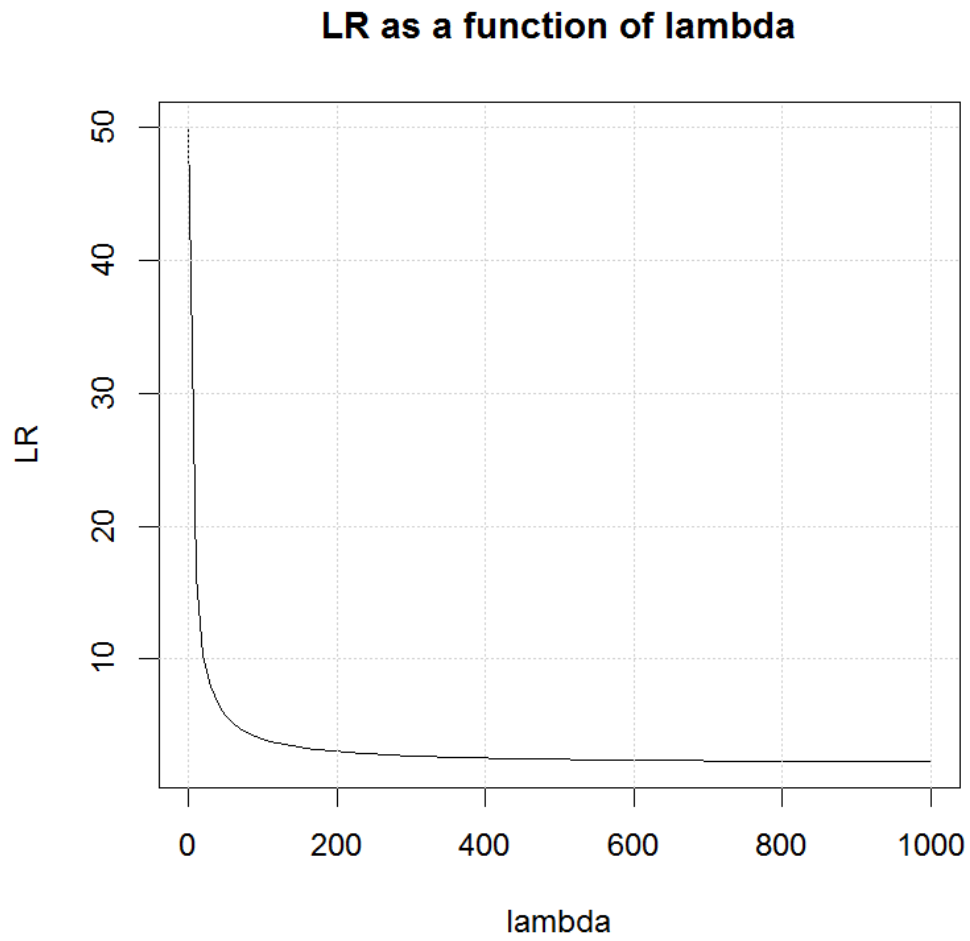Figure 4.3: Results for Exercise 4.11 e)

LR as a function of lambda

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