Exercises chapter 5

5.6.1 Expected LRs for diallelic SNPs with different allele frequency distributions

Different allele frequency distributions may yield different LR distributions. This information could be relevant when designing new DNA marker panels for kinship testing.

Study, by simulations, LR distributions for a full sibling vs unrelated case scenario for which the genotypes for one diallelic SNP is available. Compare three different allele frequency distributions; minor allele frequency, 0.1, 0.25 and 0.5. Assume the two potential siblings Sib-1 and Sib-2 will be genotyped. Let H_1: Sib-1 and Sib-2 are full siblings and H_2: Sib-1 and Sib-2 are unrelated. Mutation rates, population substructure, and other complicating factors can be disregarded.

(a) Manually define databases for each minor allele frequency given above.
(b) Define the pedigrees and generate data using 10,000 simulations. Set the seed to 1234. Compute the LR for each simulation and true hypothesis. Summarize the results as the median LR and the probability that the LR will exceed 1.
(c) Repeat (a)-(b) for each minor allele frequency.
(d) Open the frequency database containing data for 100 SNPs and repeat (b).
(e) Repeat (d) for each minor allele frequency.

5.6.2 Comparison between expected LRs for a microhaplotype marker panel and a SNP marker panel

The purpose of this exercise is to study, by simulations, LR distributions for a full sibling versus unrelated case scenario for which data are available for 40 microhaplotype loci (frequency distribution as in Fig. 5.7 above). Comparisons are made to a marker panel comprising 40 diallelic SNPs (with a 0.4/0.6 frequency distribution per marker). Let H_1: Sib-1 and Sib-2 are full siblings and H_2: Sib-1 and Sib-2 are unrelated. Frequency data are given in the online supplementary files..

(a) Create a project in Familias with the microhaplotype marker data. Define the pedigrees, set the seed to 1234 and generate data using 10,000 simulations, conditioned both on H_1 and H_2. Summarize the results as the median LR and the probability that the LR will exceed 1. *Create distribution plots and exceedance plots, i.e., plot estimates of P(LR > x|Hi), i=1, 2, for a range of x-values.
(b) Repeat (a), with the SNP marker data.
(c) Compare the results from (a) and (b).

5.6.3 *Relationship inference from a large number of SNP markers

In this exercise, we have SNP data for two individuals and we would like to estimate the degree of relationship between the two individuals. More specifically, we have genotype data for 23,742 SNPs all located on chromosome 1. In this exercise, we will focus on the segment approach, P(IBD ¼ 0) and kinship coefficient. Data are given in the online supplementary files.

(a) What is the estimated length of shared segments for a parent-child duo? Assume the total length of the markers included in the exercise is 249 cM.
(b) What is the expected P(IBD=0) and the kinship coefficient for a first cousin relationship?
(c) Estimate the total length of shared segments. Use a threshold of 7 cM and 100 SNPs to include a segment in the accumulation toward a total length.
(d) Estimate P(IBD=0) and the kinship coefficient.

5.6.4 Biogeographical ancestry prediction

Assume that we want to predict from which population Mr X originates and that we have DNA data for a single SNP marker, for which this unknown individual is homozygous for allele T. We have three different populations P1, P2 and P3. With the following allele frequencies P1 (T:0.1, C:0.9), P2 (T:0.9, C:0.1) and P3, (T:0.5, C:0.5) . Use a Bayesian approach and calculate the posterior probabilities for the hypotheses that Mr. X belongs to P1, P2 and P3. Assume HWE and a flat prior.

5.6.5 Finding LR distributions for marker panels with length based allele frequencies and marker panels with DNA sequence based allele frequencies

In Example 5.1, we compared the difference in LRs reflecting if sequence variation in STR alleles is accounted for. We considered two single markers. In this exercise, we will repeat this approach by comparing expected LRs for the 27 STR markers included in the ForenSeq marker panel. We have access to length-based allele frequencies and to sequence-based allele frequencies for a Swedish population. We would like to study how informative this panel is for a first cousin versus unrelated case. Allele frequencies are given in the online supplementary files.

(a) For each database, create a project in Familias and define persons and pedigrees.
(b) Perform simulations and estimate the median LR, mean LR, and 1%-99% intervals. Estimate exceedance probabilities (LR > 10, LR > 100, LR > 1000) for the two different allele frequency distributions. Set the seed to 1234 in the simulation window in Familias and perform 10,000 simulations for each allele database and true hypothesis. (In this exercise we ignore possible impact of linkage).
(c) *Plot LR distributions and exceedance probability distributions.

5.7.6 Estimation of IBD coefficients

In this exercise, we will estimate IBD coefficients using R.

a) Create and plot a first cousin pedigree. *Hint*: See documentation of the function cousinPed in the R library pedtools:

b) Find the kappa coefficients between the first cousins. *Hint*: See the documentation of the function kappaIBD in the R library ribd.

c) Simulate 100 equifrequent SNP markers. *Hint*: See documentation of the function markerSim in R library forrel

d) Estimate the IBD coefficients. Pedigree information is not used. *Hint*: See documentation of the appropriate function in R library forrel

e) Try different seeds and reestimate.

f) Increasing the number of markers (say to N = 10,000) will yield kappa values that fit with the theoretical values 0.75, 0.25, and 0. Verify this.

g) Show the result in the IBD triangle. *Hint*: Use the function showInTriangle.