ISFG summer school - virtual edition 2021

Pedigree analysis in R

Magnus Dehli Vigeland and Thore Egeland

Exercise set V. Relatedness inference and pedigree reconstruction

Before you start, load the packages needed in the exercises.

```
library(pedsuite)
library(pedbuildr)  # not a core package so must be loaded separately
```

If you haven't done it already, you should also download the `data` folder containing the datasets:

```
download.file("https://magnusdv.github.io/pedinr/datasets/data.zip", dest = "data.zip")
unzip("data.zip")
```

Exercise V-1 (Pairwise estimates)

This exercise uses the dataset `reconstruct-example.ped` in the `data` folder.

a) Use the following code to load and inspect the data. How many individuals are there? what is their sex? How many markers?

```
w = readPed("data/reconstruct-example.ped", locus = "snp-12")
summary(w)
```

b) Use `ibdEstimate()` to estimate kappa between individuals 2 and 4, and plot the corresponding point in the IBD triangle. How do you think they are related? How confident are you?

c) Use the code below to compute and plot 200 bootstrap simulations of the kappa coefficients.

```
ibdBootstrap(w, ids = c(2,4), param = "kappa", N = 200)
```

Comment on the uncertainty of the estimate in b).

d) Now estimate and plot the kappa coefficients between all pairs of individuals in the dataset. For each pair, explain what the kappa estimate tells us about their pedigree relationship. Which relationships are you most confident about?

Exercise V-2 (Pedigree reconstruction)

Use the following **pedbuildr** command to reconstruct the pedigree from the previous exercise.

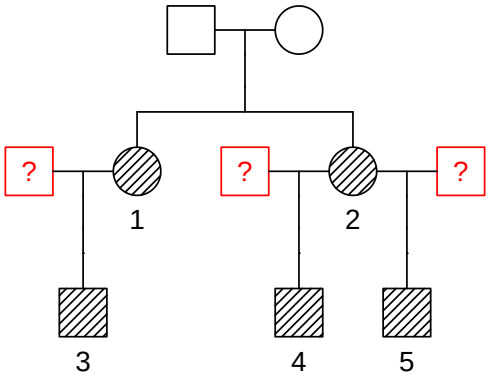
```
# Reconstruct
r = reconstruct(w, inferPO = TRUE, linearInb = FALSE)

# Plot the most likely pedigrees
plot(r)
```

a) Explain the meaning of each argument in the function call.
b) Study the verbose output of the function. How many pedigrees did it consider?
c) What is the most likely pedigree? How much more likely is it compared with the runner-up?

Exercise V-3 (A question about fathers)

The following is based on a true case from Australia. Genotypes are available from two sisters and their children. The first sister has one child, the other has two children. The question we must answer is: *Do any of the children have the same father?*

The data are given in the files `reconstruct-fathers.ped` and `reconstruct-fathers.freq`.



a) Load the data with the commands below, and use `summary()` to inspect the data. Check that the labels and sexes match the figure. How many markers are used?

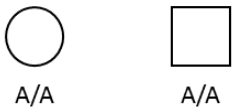
```
x = readPed("data/reconstruct-fathers.ped")
x = setFreqDatabase(x, "data/reconstruct-fathers.freq")
summary(x)
```

b) Use **pedbuildr** to perform a pedigree reconstruction on the data set. *Hint*: Use undisputed parts of the family to restrict the search space. For example, the parameters `knownPO` and `noChildren` might be useful. You should also add `linearInb = F` to exclude parent-child incest etc.

c) Plot the six most likely pedigrees and describe the paternity constellations. What is your conclusion?

Exercise V-4 (Bonus exercise for the mathematically inclined: ML-estimation by hand!)

This exercise walks you through the computations of a maximum likelihood estimation of the relationship between two non-inbred individuals. To enable hand calculation, the data is the simplest possible: Genotypes from a single marker, for which both individuals are homozygous A/A. Denote by p the population frequency of the A allele, and assume $0 < p < 1$.



a) What do you think is the most likely relationship given this data?
Recall that maximum likelihood estimation of pairwise relatedness works by finding the value of $k = (k_0, k_1, k_2)$ that maximises the likelihood function

$$\begin{aligned} L(k) &= P(data \mid k) && (1)\\ &= P(data \mid UN) \cdot k_0 + P(data \mid PO) \cdot k_1 + P(data \mid MZ) \cdot k_2, && (2) \end{aligned}$$

where the *data* are the observed genotypes, and UN, PO and MZ denote the relationships of unrelated, parent-child and MZ twins, repsectively.

b) Show that the likelihood function in this case becomes

$$L(k) = P(AA, AA \mid k) = p^4 k_0 + p^3 k_1 + p^2 k_2. \quad (3)$$

c) Explain that for optimisation purposes you can get rid of a factor p^2. Furthermore, use the relation $k_0 + k_1 + k_2 = 1$ to eliminate k_1, giving the simpler function

$$L_1(k_0, k_2) = (p^2 - p)k_0 + p + (1 - p)k_2. \quad (4)$$

d) Remove another factor, $(1 - p)$, and conclude that all we have to do is to maximize the function

$$L_2(k_0, k_2) = k_2 - pk_0 + C, \quad (5)$$

where C is a constant.

e) Which point (k_0, k_2) in the IBD triangle gives the highest value of the function L_2 obtained in the previous step? What is the estimated relationship?