Weight of DNA evidence using the forensim package

Thore Egeland

Hinda Haned

June 2010

Contents

1 The forensim package: overview

1.1 Available documentation

forensim is an R-package hosted by R-forge dedicated to facilitate the interpretation of forensic DNA mixtures. It also provides simulation tools made to mimick data from case work.

A detailed description of forensim is given in the package tutorial, available from: http://forensim.r-forge.r-project.org/. prepared specifically for potential forensim users who are unfamilliar with R. The present document serves to

- introduce the basic statistical calculations of forensim,
- provided excercises and solutions for a course setting,
- provide examples to verify correct usage and answers. This is mostly done by means of the solution to the mentioned excercises.

1.2 Statistical methods. A worked example

Forensim provides a variety of methods dedicated to evaluating the weight of DNA evidence [1]. Below we focus on the `LR` function for the calculation of likelihood ratios. The `LR` function implements the general formula of Curran et al. for forensic DNA mixtures interpretation [2].

An example Consider the following genetic profiles from a rape case in Hong Kong [3]:

Locus	Mixture	Victim	Suspect	Frequency
D3S1358	14		14	0.033
	15	15		0.331
	17		17	0.239
	18	18		0.056

Table 1: Alleles from a DNA stain from a rape case in Hong Kong

Locus D3S1358 shows 4 distinct alleles (14, 15, 17 and 18). The number of contributors to the mixed sample is taken to be 2.

Scenario 1 The following hypotheses are tested:

- Prosecution hypothesis H_P: Contributors were the victim and the suspect.
- Defence hypothesis H_D: Contributors were 2 unknown people.

Before we start, remember to load the package:

```
> library(forensim)


   ### forensim 1.1.9 is loaded ###
```

First, the genotypes are assigned to the victim and the suspect:

```
> victim <- "15/18"
> suspect <- "14/17"
```

The likelihood ratio is computed using the `LR` function: Here is a useful extract of this function's help page:

- `stain`: a vector giving the set of (distinct) alleles present in the DNA stain
- `freq`: vector of the corresponding allele frequencies in the global population
- `xp`: the number of unknown contributors to the stain under the prosecution hypothesis Hp. Default is 0.
- `xd`: the number of unknown contributors to the stain under the defence hypothesis Hd. Default is 0.
- `Tp`: a vector of strings where each string contains two alleles separated by '/', corresponding to one known contributor under the prosecution hypothesis Hp. The length of the vector equals the number of known contributors. Default is NULL.
- `Vp`: a vector of strings where each string contains two alleles separated by '/', corresponding to one known non-contributor under the prosecution hypothesis Hp. The length of the vector equals the number of known non-contributors. Default is NULL.
- `Td`: a vector of strings where each string contains two alleles separated by '/', corresponding to one known contributor under the defence hypothesis Hd. The length of the vector equals the number of known contributors. Default is NULL.
- `Vd`: a vector of strings where each string contains two alleles separated by '/', corresponding to one known non-contributor under the defence hypothesis Hd. The length of the vector equals the number of known non-contributors. Default is NULL.
- `theta`: a float in [0,1[. theta is equivalent to Wright's Fst. In case of population subdivision, it allows a correction of the allele frequencies in the subpopulation of interest

The LR is obtained as follows

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = NULL, Vd = c(victim,
+         suspect), xd = 2, theta = 0)


NOTE: THIS PACKAGE IS NOW OBSOLETE.

  The R-Genetics project has developed an set of enhanced genetics
  packages to replace 'genetics'. Please visit the project homepage
  at http://rgenetics.org for informtion.

[1] 285
```

The mixture profile is 285 times more likely if it came from the suspect and the victim than if it came from two unknown unrelated individuals.
Note that as long as theta=0, there is no need to be specify the non-contributing individuals, so the same figure is prodcued with Vd=NULL.

Scenario 2 The following hypotheses are tested:
Prosecution hypo4thesis H_P: Contributors were the victim and the suspect.
Defence hypothesis H_D: Contributors were the victim and one unknown.

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = victim,
+     Vd = suspect, xd = 1, theta = 0)
```

```
[1] 63.4
```

The mixture profile is 63 times more likely if it came from the suspect than if it came from an unrelated individual.

2 Exercises

Some of the problems below are theoretical in the sense that forensim is not used, rather calculation by hand are requested. These excercises may be skipped for those exclusively interested in practising forensim.

2.1 Excercise 1. Likelihood ratios and theta values

Note that in the previous examples, the `theta` argument does not appear in the `LR` function. This means that the argument is set to its default value, which is 0. The problems below extend on scenario 2 of the above example by addressing θ corrections.

1. Change the value of the `theta` argument from 0 to 0.03 and repeat the calculation.

2. Calculate the LR for different values of `theta` taken in the interval [0,0.03].

    - *Tip 1*: use the `seq` function to create a sequence of values for the `theta` argument.
    - *Tip 2*: use the *sapply* function to compute the values of the LR for different values of theta. To get help, type: help('sapply').

3. Represent the obtained results in a plot (use function `plot`).

2.2 Excercise 2. Theoretical continuation of Excercise 1

1. Derive the formulae corresponding to Scenarios 1 and 2 of the worked example (Hong-Kong case). Confirm that the figures obtanined by forensim are correct.

2. Repeat the above problem with θ-correction for scenario 2 ($\theta = 0.03$).

2.3 Excercise 3. LR-calculations for mixtures

The purpose of this exercise is to demonstrate various approaches to LR calculations for a mixture case. The data comes from a proficiency test arranged by GEDNAP `http://gednap.de/`. For simplicity only three markers are considered. There is a mixture (stain), the data is summarised in Table˜2 and a reference sample is shown in Table˜3.

Locus	Allele
D3S1358	15
D3S1358	16
D3S1358	17
vWA	15
vWA	16
vWA	18
FGA	20
FGA	21
FGA	22
FGA	24
FGA	26

Table 2: The crime scene profile at three STR loci.

Locus	Allele
D3S1358	15/17
vWA	16/18
FGA	20/26

Table 3: The reference sample B

The hypotheses are

- H_P: B and two unknown individuals contributed to the stain
- H_D: Three unknown people contributed to the stain

Calculate the likelihood ratios to weight hypotheses H_P and H_D when $\theta = 0$. For simplicity we assume all allele frequencies to be 0.1.

2.4 Excercise 4. LR: standard and for drop in and out

This example extends on Section 4.4 of [4]. The hypotheses are the usual ones:

- H_P: The DNA came from the suspect.
- H_D: The DNA came from a random man.

Throughout A and B denote alleles with relative frequencies $p_A = 0.2$ and $p_B = 0.1$, and we assume $\theta = 0$.

1. We first consider a standard case with data AB for the suspect and the stain. Derive the formula for the LR and use R to provide the numeric answer. Confirm the above calculation using the LR function of forensim.

2. Repeat the above problem whith data AA for suspect and the stain.

3. Assume markers $1, 2, \cdots, 5$ are as 1 above. Markers 6,7,8,9 are as for 2 above. Calculate the LR for these 9 markers by using the formulae derived above.

4. For the tenth marker the suspect is A and the stain AB. What's the LR for this marker? What's the LR based on all 10 markers?

5. Consider the above problem once assuming that there is a probability D that an allele drops out. According to [4]

$$LR_{10} \approx \frac{D}{(1+D)p_A^2 + 2p_A(1-p_A)D} \tag{1}$$

   Let $D = 0.1$. Use R to find LR_{10} and the LR based on all markers ($LR_{1,10}$). Comment on the answer.

6. Plot LR_{10} as a function of D.

3 Solutions to the excercises

3.1 Excercise 1

1. For a single value, theta=0.03 we find:

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = victim,
+     Vd = suspect, xd = 1, theta = 0.03)

[1] 37.6
```

2. Define a variable theta, taking different values in the [0,1] interval:

```
> theta <- seq(0, 0.03, by = 0.001)
> theta

 [1] 0.000 0.001 0.002 0.003 0.004 0.005 0.006 0.007 0.008 0.009 0.010 0.011
[13] 0.012 0.013 0.014 0.015 0.016 0.017 0.018 0.019 0.020 0.021 0.022 0.023
[25] 0.024 0.025 0.026 0.027 0.028 0.029 0.030
```

   To replicate the calculations for different values of theta, we use the `sapply` function.

```
> sapply(theta, function(i) LR(stain = c(14, 15, 17, 18), freq = c(0.033,
+     0.331, 0.239, 0.056), xp = 0, Tp = c(victim, suspect), Vp = NULL,
+     Td = victim, Vd = suspect, xd = 1, theta = i))

 [1] 63.40 61.83 60.34 58.94 57.61 56.35 55.15 54.01 52.92 51.89 50.90 49.95
[13] 49.05 48.18 47.35 46.56 45.79 45.06 44.35 43.67 43.02 42.39 41.78 41.19
[25] 40.63 40.08 39.55 39.04 38.54 38.06 37.60
```

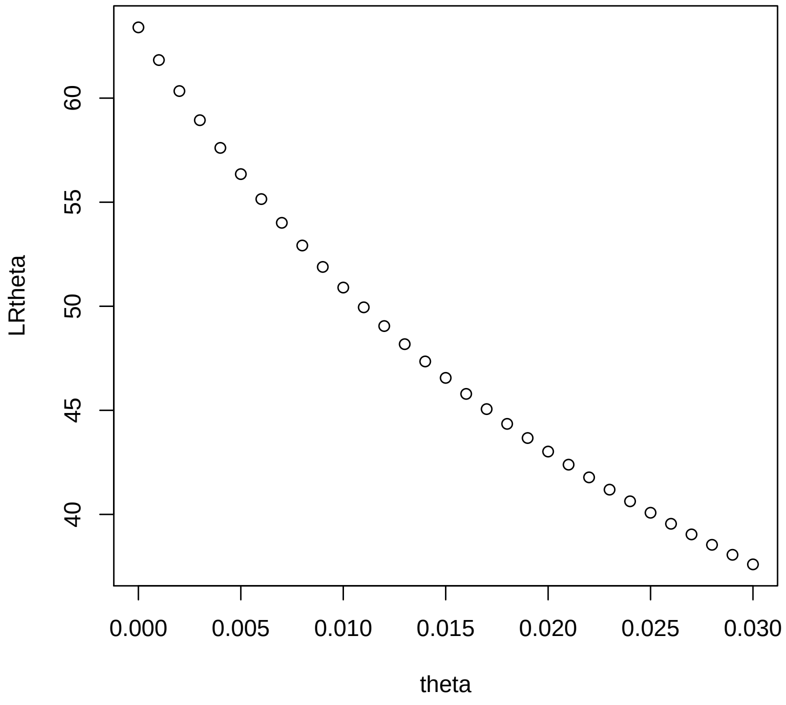
The above command calculates the LR for each value `i` of `theta`.

3. To plot these results, we need first to save them to an object:

```
> LRtheta <- sapply(theta, function(i) LR(stain = c(14, 15, 17,
+     18), freq = c(0.033, 0.331, 0.239, 0.056), xp = 0, Tp = c(victim,
+     suspect), Vp = NULL, Td = victim, Vd = suspect, xd = 1, theta = i))
```

The plot is produced by

```
> plot(theta, LRtheta)
```



3.2 Excercise 2

1. For scenario 1

$$LR = \frac{1}{24p_{14}p_{15}p_{17}p_{18}} = \frac{1}{24 \cdot 0.033 \cdot 0.331 \cdot 0.239 \cdot 0.056}. \tag{2}$$

   The numerator is obvious. The denominator can be be obtained by realising that both individuals must be heterozygote and that there are 6 possible combinations, each having probability $4p_{14}p_{15}p_{17}p_{18}$ since (i) Hardy-Weinberrg Equilibrium is assumed to hold and (ii) the individuals are unrelated.

   This can be calculated in R as

```
> 1/(24 * 0.033 * 0.331 * 0.239 * 0.056)
```

```
[1] 285.0105
```

   For scenario 2

$$LR = \frac{1}{2p_{14}p_{17}} \tag{3}$$

   since the suspect must have genotype 14,17.

   This can be calculated in R as

```
> 1/(2 * 0.033 * 0.239)
```

```
[1] 63.39546
```

2. Consider first scenario 1. Let $A = 14, B = 15, C = 17, D = 18$. Then the modification of Equation˜3 to account for θ-corrections becomes

$$LR = \frac{(1+3\theta)(1+4\theta)}{2(\theta + (1-\theta)p_{14})(\theta + (1-\theta)p_{17})}. \tag{4}$$

```
> (1 + 3 * 0.03) * (1 + 4 * 0.03)/(2 * (0.03 + (1 - 0.03) * 0.033) *
+     (0.03 + (1 - 0.03) * 0.239))
```

```
[1] 37.59529
```

   This confirms the forensim value:

```
> LR(stain = c(14, 15, 17, 18), freq = c(0.033, 0.331, 0.239, 0.056),
+     xp = 0, Tp = c(victim, suspect), Vp = NULL, Td = victim,
+     Vd = suspect, xd = 1, theta = 0.03)
```

```
[1] 37.6
```

3.3 Excercise 3

First, enter the stain profile for each available locus:

```
> stainD3 <- c(15, 16, 17)
> stainv <- c(15, 16, 18)
> stainFGA <- c(20, 21, 22, 24, 26)
```

Second, enter the suspect profile for each available locus:

```
> suspectD3 <- "15/17"
> suspectv <- "16/18"
> suspectFGA <- "20/26"
```

Last, the likelihood ratio:

```
> LRD3 <- LR(stain = stainD3, freq = rep(0.1, 3), xp = 2, Tp = c(suspectD3),
+     Vp = NULL, Td = NULL, Vd = suspectD3, xd = 3, theta = 0)
> LRD3


[1] 12.04


> LRDv <- LR(stain = stainv, freq = rep(0.1, 3), xp = 2, Tp = c(suspectv),
+     Vp = NULL, Td = NULL, Vd = suspectv, xd = 3, theta = 0)
> LRDv


[1] 12.04


> LRDFGA <- LR(stain = stainFGA, freq = rep(0.1, 5), xp = 2, Tp = c(suspectFGA),
+     Vp = NULL, Td = NULL, Vd = suspectFGA, xd = 3, theta = 0)
> LRDFGA


[1] 4.667
```

The overall likelihood ratio is obtained by multiplying the above likelihood ratios:

```
> LRD3 * LRDv * LRDFGA


[1] 676.5358
```

3.4 Excercise 4

1. Note first that $P(data|H_P) = 1$. Next

   $$P(data|H_D) = P(\text{culprit is AB}) = 2p_Ap_B$$

   provided Hardy-Weinbererg Equilibrium holds. First some parameter values are assigned.

   ```
   > D <- 0.01
   > pA <- 0.2
   > pB <- 0.1
   ```

   A direct calculation in R gives

   ```
   > 1/(2 * pA * pB)

   [1] 25
   ```

   The LR function of forensim gives

   ```
   > LR(stain = c("A", "B"), freq = c(0.2, 0.1), xp = 0, Tp = "A/B",
   +     Vp = NULL, Td = NULL, Vd = "A/B", xd = 1, theta = 0)

   [1] 25
   ```

2. A similar argument gives $LR = 1/p_A^2$ which evaluates to 25. Furthermore, using forensim we find

   ```
   > LR(stain = c("A"), freq = c(0.2), xp = 0, Tp = "A/A", Vp = NULL,
   +     Td = NULL, Vd = "A/A", xd = 1, theta = 0)

   [1] 25
   ```

3.
   ```
   > 25^9

   [1] 3.814697e+12
   ```

4. The likelihood ratio is 0 for the marker and therefore also the overall LR is 0.

5. With drop-out probability of 0.01 we find

   ```
   > D/((1 + D) * pA^2 + 2 * pA * (1 - pA) * D)

   [1] 0.2293578
   ```

   The LR based on all markers becomes

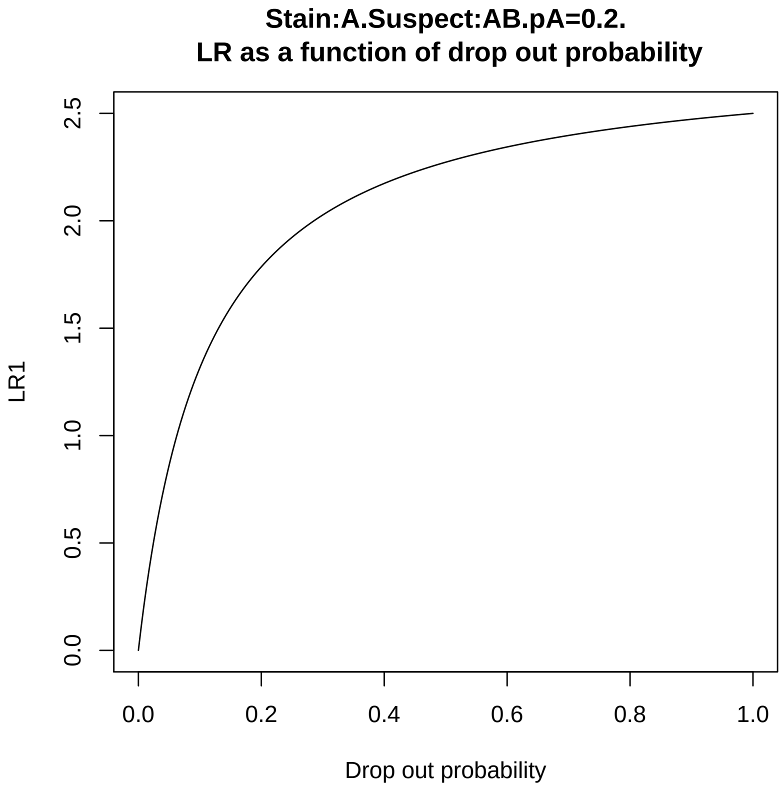
   ```
   > 25^9 * 0.2293578
   ```

```
[1] 874930572510
```

6. There are several ways to plot. Here's one:

```
> D = seq(0, 1, length = 1000)
> pA = 0.2
> LR1 = D/((1 + D) * pA^2 + D * 2 * pA * (1 - pA))
```

```
> D = seq(0, 1, length = 1000)
> pA = 0.2
> LR1 = D/((1 + D) * pA^2 + D * 2 * pA * (1 - pA))
> plot(D, LR1, type = "l", xlab = "Drop out probability", ylab = "LR1")
> title("Stain:A.Suspect:AB.pA=0.2.\n LR as a function of drop out probability")
```



References

[1] H.~Haned. Forensim: an open source initiative for the evaluation of statistical methods in forensic genetics. *Forensic Sci. Int. Genetics*, 2010.

[2] J.~Curran, J.~Buckleton, and C.~M. Triggs. What is the magnitude of the subpopulation effect? *Forensic Science International*, 135:1–8, 2003.

[3] W.~K. Hu and W.~K. Fung. Interpreting dna mixtures with the presence of relatives. *International Journal of Legal Medicine*, 117:39–45, 2003.

[4] P.~Gill and J.~Buckleton. A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number. *Forensic Science International: Genetics*, 4(4):221–227, 2010.