The statistical evaluation of DNA crime stains in R

Miriam Marušiaková

Department of Statistics, Charles University, Prague
Institute of Computer Science, CBI, Academy of Sciences of CR

UseR! 2008, Dortmund
Introduction

Single Crime Scene
   Independence
   Population substructure
   Relatedness
   R package forensic

Mixed stain
   Independence
   Population substructure
   People versus Simpson
   More general situation

Conclusion
Outline

Introduction

Single Crime Scene
  Independence
  Population substructure
  Relatedness
  R package forensic

Mixed stain
  Independence
  Population substructure
  People versus Simpson
  More general situation

Conclusion
Introduction

Chromosomal locations of RFLP markers used in DNA profiling

D1S7  D2S44  D4S139

1  2  3  4  5

6  7  8  9  10

D10S28  D17S79

11  12  13  14  15  16  17  18

Human Female Karyotype
Single crime scene stain

- Blood stain at the crime scene
- Believed it was left by offender
- Suspect arrested for reasons unconnected with his DNA profile
- Crime sample, suspect sample

Hypothesis

- \( H_p \) (prosecution): The suspect left the crime stain.
- \( H_d \) (defense): Some other person left the crime stain.

Notation

- DNA typing results \( E = \{G_C, G_S\} \)
- non-DNA evidence \( I \)
Evidence interpretation

- Prior odds
  \[
  \frac{Pr(H_p | I)}{Pr(H_d | I)}
  \]

- Posterior odds
  \[
  \frac{Pr(H_p | E, I)}{Pr(H_d | E, I)}
  \]

- Bayes’ theorem
  \[
  \frac{Pr(H_p | E, I)}{Pr(H_d | E, I)} = \frac{Pr(E | H_p, I)}{Pr(E | H_d, I)} \times \frac{Pr(H_p | I)}{Pr(H_d | I)}
  \]
  \[
  LR
  \]

Evidence interpretation

\[
LR = \frac{Pr(E|H_p, I)}{Pr(E|H_d, I)} = \frac{Pr(G_S, G_C|H_p, I)}{Pr(G_S, G_C|H_d, I)} \\
= \frac{Pr(G_C|G_S, H_p, I)}{Pr(G_C|G_S, H_d, I)} \times \frac{Pr(G_S|H_p, I)}{Pr(G_S|H_d, I)} \\
= \frac{1}{Pr(G_C|G_S, H_d, I)} \times 1 \\
= \frac{1}{Pr(G_C|H_d, I)} \quad \text{if independence assumed}
\]
Errors and fallacies

Statement
The probability of observing this type
if the blood came from someone other than suspect
is 1 in 100.

\[ Pr(G_C|H_d, I) = 1/100 \]

Common error
The probability that the blood came from someone else is 1 in 100.

\[ Pr(H_d|G_C, I) = 1/100 \]

There is a 99% probability that it came from the suspect.
Errors and fallacies (cont.’d)

Statement

The evidence is 100 times more probable if the suspect left the crime stain than if some unknown person left it.

\[
\frac{1}{Pr(G_C|H_d, I)} = 100
\]

Common error

It is 100 times more probable that the suspect left the crime stain than some unknown person.

\[
\frac{Pr(H_p|G_C, I)}{Pr(H_d|G_C, I)} = 100
\]
Outline

Introduction

Single Crime Scene
  Independence
  Population substructure
  Relatedness
  R package forensic

Mixed stain
  Independence
  Population substructure
  People versus Simpson
  More general situation

Conclusion
Outline

Introduction

Single Crime Scene

Independence
Population substructure
Relatedness
R package forensic

Mixed stain

Independence
Population substructure
People versus Simpson
More general situation

Conclusion
Product rule

Model of an ideal population

- Infinite size
- Random mating
- Model reliable for most real-world problems

Hardy-Weinberg equilibrium

\[
Pr(G = A_iA_i) = p_i^2 \\
Pr(G = A_iA_j) = 2p_ip_j, \quad i \neq j
\]

Likelihood ratio

\[
LR = \frac{1}{Pr(G_C|G_S, H_d, I)} = \frac{1}{Pr(G_C|H_d, I)} = \frac{1}{p_i^2} \left( \frac{1}{2p_ip_j} \right)
\]
Outline

Introduction

Single Crime Scene
  Independence
  Population substructure
  Relatedness
  R package forensic

Mixed stain
  Independence
  Population substructure
  People versus Simpson
  More general situation

Conclusion
Population substructure

- $F$ - measure of uncertainty about allele proportions in the population of suspects
- Genetic interpretation of $F$ (Wright, 1951)
- How to estimate $F$ (Weir and Cockerham, 1984)
- Recommendations (National Research Council, 1996)

$F = 0.01$ large subpopulations (USA)
$F = 0.03$ small isolated subpopulations

- Match probabilities (Balding and Nichols, 1994)

$$P(G_C = A_iA_i|G_S = A_iA_i) = \frac{[2F + (1 - F)p_i] [3F + (1 - F)p_i]}{(1 + F)(1 + 2F)}$$

$$P(G_C = A_iA_j|G_S = A_iA_j) = \frac{2 [F + (1 - F)p_i] [F + (1 - F)p_j]}{(1 + F)(1 + 2F)}$$
Effects of $F$ corrections

Likelihood ratio - some numerical values

- **Heterozygotes** $A_iA_j$, $p_i = p_j = p$

<table>
<thead>
<tr>
<th>$F$</th>
<th>$p = 0.01$</th>
<th>$p = 0.05$</th>
<th>$p = 0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F = 0$</td>
<td>5000</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>$F = 0.001$</td>
<td>4152</td>
<td>193</td>
<td>49</td>
</tr>
<tr>
<td>$F = 0.01$</td>
<td>1295</td>
<td>145</td>
<td>43</td>
</tr>
<tr>
<td>$F = 0.03$</td>
<td>346</td>
<td>89</td>
<td>34</td>
</tr>
</tbody>
</table>

- **Homozygotes** $A_iA_i$, $p_i = p$

<table>
<thead>
<tr>
<th>$F$</th>
<th>$p = 0.01$</th>
<th>$p = 0.05$</th>
<th>$p = 0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F = 0$</td>
<td>10000</td>
<td>400</td>
<td>100</td>
</tr>
<tr>
<td>$F = 0.001$</td>
<td>6439</td>
<td>364</td>
<td>96</td>
</tr>
<tr>
<td>$F = 0.01$</td>
<td>863</td>
<td>186</td>
<td>67</td>
</tr>
<tr>
<td>$F = 0.03$</td>
<td>157</td>
<td>73</td>
<td>37</td>
</tr>
</tbody>
</table>
Outline

Introduction

Single Crime Scene
  Independence
  Population substructure
  Relatedness
  R package forensic

Mixed stain
  Independence
  Population substructure
  People versus Simpson
  More general situation

Conclusion
Inbreeding

- Individuals with common ancestors - related
- Their children - inbred
- Alleles are ibd (identical by descent) - copies of the same allele

Example

Alleles $h_1, h_2$ transmitted from parent $H$ to $X, Y$ who transmit $a, b$ to $I$

\[
Pr(h_1 \text{ is ibd to } h_2) = Pr(h_1 \equiv h_2) = 0.5
\]
\[
Pr(a \equiv b) = Pr(a \equiv h_1, b \equiv h_2 | h_1 \equiv h_2)Pr(h_1 \equiv h_2) = 0.5^3 = 0.125
\]
Match probabilities for close relatives

Balding and Nichols (1994)

\[ Pr(G_C|G_S, H_d, I) = \begin{cases} 
\frac{(k_0 p_i^4 + k_1 p_i^3 + k_2 p_i^2)}{p_i^2} \\
\frac{4k_0 p_i^2 p_j^2 + k_1 p_i p_j (p_i + p_j) + 2k_2 p_i p_j}{2p_i p_j} 
\end{cases} \]

Kinship coefficients

<table>
<thead>
<tr>
<th>Relationship</th>
<th>( k_0 )</th>
<th>( k_1 )</th>
<th>( k_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent - child</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Siblings</td>
<td>1/4</td>
<td>1/2</td>
<td>1/4</td>
</tr>
<tr>
<td>Grandparent - grandchild</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>Uncle - nephew</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>Cousins</td>
<td>3/4</td>
<td>1/4</td>
<td>0</td>
</tr>
<tr>
<td>Unrelated</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\( k_i \) - probability that two persons will share \( i \) alleles ibd
\( i = 0, 1, 2 \)
Outline

Introduction

Single Crime Scene
- Independence
- Population substructure
- Relatedness
  R package forensic

Mixed stain
- Independence
- Population substructure
- People versus Simpson
- More general situation

Conclusion
R function Pmatch

Usage

Pmatch(prob, k = c(1, 0, 0), theta = 0)

Example

> p <- c(0.057, 0.160, 0.182, 0, 0.024, 0.122)
> Pmatch(p)

$prob

    locus 1  locus 2  locus 3
allele 1  0.057  0.182  0.024
allele 2  0.160  0.000  0.122

$match

    locus 1  locus 2  locus 3
[1,]  0.01824  0.033124  0.005856

$total_match

[1] 3.538088e-06
Outline

Introduction

Single Crime Scene
- Independence
- Population substructure
- Relatedness
- R package forensic

Mixed stain
- Independence
- Population substructure
- People versus Simpson
- More general situation

Conclusion
Mixtures

- Prosecution and defense hypothesis
  - \( H_p \): Contributors were the victim and the suspect.
  - \( H_d \): Contributors were the victim and an unknown person.

- Likelihood ratio for the mixture

\[
L_R = \frac{Pr(E, G_V, G_S|H_p, I)}{Pr(E, G_V, G_S|H_d, I)}
= \frac{Pr(E|G_V, G_S, H_p, I)}{Pr(E|G_V, G_S, H_d, I)} \times \frac{Pr(G_V, G_S|H_p, I)}{Pr(G_V, G_S|H_d, I)}
= \frac{1}{Pr(E|G_V, G_S, H_d, I)}
= \frac{1}{Pr(E|G_V, H_d, I)}
\]

(if independence assumed)
Outline

Introduction

Single Crime Scene
  Independence
  Population substructure
  Relatedness
  R package forensic

Mixed stain
  Independence
  Population substructure
  People versus Simpson
  More general situation

Conclusion
Match probability (independence assumption)
Weir et al (1997)

\[ P_x(U, C) = (T_0)^{2x} - \sum_{i \in U} (T_{1i})^{2x} + \sum_{i,j \in U : i < j} (T_{2ij})^{2x} - \ldots \]

\[
T_0 = \sum_{l \in C} p_l \\
T_{1i} = \sum_{l \in C \setminus \{i\}} p_l, \quad i \in U \\
T_{2ij} = \sum_{l \in C \setminus \{i,j\}} p_l, \quad i,j \in U, \quad i < j
\]

\(U\) - set of alleles from the crime sample \(C\) not carried by known contributors
\(x\) - number of unknown contributors

R

- `Pвид.ind(alleles, prob, x, u = NULL)`
- `LR.ind(alleles, prob, x1, x2, u1 = NULL, u2 = NULL)`
Outline

Introduction

Single Crime Scene
  Independence
  Population substructure
  Relatedness
  R package forensic

Mixed stain
  Independence
  Population substructure
  People versus Simpson
  More general situation

Conclusion
Match probability for structured population

Assumption
All the involved people come from the same subpopulation with parameter $F$


$$MP = \sum_{r_1=0}^{r} \sum_{r_2=0}^{r-r_1} \cdots \sum_{r_{c-1}=0}^{r-r_1-\cdots-r_{c-2}} \frac{(2n_U)! \prod_{i=1}^{c} \prod_{j=t_i+u_i+v_i}^{t_i+u_i+v_i-1} [(1 - F)p_i + jF]}{\prod_{i=1}^{c} u_i! \prod_{j=2n_T+2n_U+2n_V}^{2n_T+2n_U+2n_v-1} [(1 - F) + jF]}$$

R

- `Pevid.gen(alleles, prob, x, T = NULL, V = NULL, theta = 0)`

$T, V$ - genotypes of known contributors, known non-contributors
Outline

Introduction

Single Crime Scene
  Independence
  Population substructure
  Relatedness
  R package forensic

Mixed stain
  Independence
  Population substructure
  People versus Simpson
  More general situation

Conclusion
People versus Simpson (Los Angeles County Case)

DNA evidence

- Material at the crime scene - alleles $a$, $b$, $c$ (locus D2S44)
- Suspect’s genotype $ab$
- Victim’s genotype $ac$


$H_p$: Contributors were the victim, suspect and $m$ unknowns.
$H_d$: Contributors were $n$ unknowns.

R

- $a = \text{c(’a’, ’b’, ’c’)}$
- $p = \text{c(0.0316, 0.0842, 0.0926)}$
- $\text{suspect <- ’a/b’, victim <- ’a/c’}$
## Likelihood ratios for the Simpson case

<table>
<thead>
<tr>
<th></th>
<th>$n = 2$</th>
<th>$n = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Defense</strong></td>
<td>$F = 0$</td>
<td>$F = 0$</td>
</tr>
<tr>
<td><strong>Prosecution</strong></td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$m = 0$</td>
<td>1623 739 276</td>
<td>21606 5853 1150</td>
</tr>
<tr>
<td>$m = 1$</td>
<td>70 44 26</td>
<td>938 345 107</td>
</tr>
</tbody>
</table>

### Prosecution proposition ($F = 0$)

- `Pевид.ind(alleles = a, prob = p, x = m)`
- `Pевид.gen(alleles = a, prob = p, x = m, T = c('victim', 'suspect'))`

### Defense proposition ($F = 0$)

- `Pевид.ind(alleles = a, prob = p, x = n, u = c('a','b','c'))`
- `Pевид.gen(alleles = a, prob = p, x = n)`
Outline

Introduction

Single Crime Scene
- Independence
- Population substructure
- Relatedness
- R package forensic

Mixed stain
- Independence
- Population substructure
- People versus Simpson

More general situation

Conclusion
More general situation

Contributors from different ethnic groups
Fung and Hu (2001)

Pevid.ind, LR.ind
(independence within and between ethnic groups)

Presence of related people
Hu and Fung (2003)

Example
Suspect not typed, his relative is tested
Two related people among unknown contributors

Pevid.rel
Outline

Introduction

Single Crime Scene
  Independence
  Population substructure
  Relatedness
  R package forensic

Mixed stain
  Independence
  Population substructure
  People versus Simpson
  More general situation

Conclusion
Conclusion

Single crime stain

- Suspect and offender are unrelated
- are members of the same subpopulation
- are close relatives

Mixed crime stain

- Contributors unrelated
- members of the same subpopulation
- may be related
- from different ethnic groups
Thank you for your attention!

Acknowledgement

The work was supported by GAČR 201/05/H007 and the project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic.
**Important publications**

**Balding and Nichols (1994)**
DNA profile match probability calculation
Forensic Science International 64, p. 125-140

**Weir et al (1997)**
Interpreting DNA mixtures
Journal of Forensic Sciences 42, p. 213-222

**Fung and Hu (2000)**
Interpreting forensic DNA mixtures: allowing for uncertainty in population substructure and dependence
Journal of Royal Statistical Society A 163, p. 241-254

**Zoubková, Supervisor Zvárová (2004)**
Master thesis (in Czech)
Charles University, Prague
Software

R DEVELOPMENT CORE TEAM (2006)
R: A language and environment for statistical computing.
R Foundation for Statistical Computing, Vienna, Austria

MARUŠIAKOVÁ, M (2007)
forensic: Statistical methods in forensic genetics
R package version 0.2