Mixture Statistics

Florida Statewide Training Meeting

Indian Rocks Beach, FL
May 12-13, 2008

Dr. John M. Butler
National Institute of Standards and Technology

john.butler@nist.gov

German Type A,B, and C mixture classifications

- **Type A**, where major/minor contributors cannot be deduced, require stats
  - LR
  - RMNE
- **Type B** enables major contributor to be deduced
  - RMP (which is 1/LR)
- **Type C** no stats should be attempted because of the possibility of failure to account for allele dropout due to stochastic effects with low level DNA samples

Statistical Interpretation of DNA Mixtures


CROATIAN MEDICAL JOURNAL 42(3):244-246, 2001

CMJ FORENSIC SCIENCES

Interpretation of Complex Forensic DNA Mixtures

Carll Ladd, Henry C. Lee, Nicholas Yang, Frederick R. Bieber[1]

[1]State of Connecticut Department of Public Safety, Division of Scientific Services, Forensic Science Laboratory, Conn; and [2]Departments of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Mass, USA

Ladd et al. 2001. *Croatian Medical Journal* 43(3): 244-246

1. **Qualitative statement ('..cannot exclude..')**
2. **Interpret as single source from peak height differences, differential extraction, etc. and calculate random match probability (RMP)**
3. **Calculate probability of exclusion (CPE)**
4. **Calculate likelihood ratio (LR)**

Random Man Not Excluded (RMNE)

- = Probability of Exclusion (PE)
- John Buckleton (*Forensic DNA Evidence Interpretation*, p. 222) quotes Laszlo Szabo of Tasmania Forensic Science Laboratory: "Intuitively, RMNE is easier to explain to a jury and express in reports than the likelihood ratio, and is probably closer to what the court wants—e.g., the suspect matches the mixture, but what if this is the wrong person– then what is the probability that someone else in the population would also match the mixture (i.e., not be excluded as a contributor)."
- Buckleton (*Forensic DNA Evidence Interpretation*, p. 222) also quotes Bruce Weir: that exclusion probabilities "often rob the items of probative value"

Probability of Exclusion (RMNE)

- **Advantages**
  - Does not require an assumption of the number of contributors to a mixture
  - Easier to explain in court
- **Disadvantages**
  - Weaker use of the available information (robs the evidence of its true probative power because this approach does not consider the suspect's genotype)
  - Likelihood ratio approaches are developed within a consistent logical framework

John Buckleton, *Forensic DNA Evidence Interpretation*, p. 223

RMNE (CPE)

- Statements from DAB Recommendations on Statistics (FDT2e, p. 617)
- CPE provides a calculation of the estimated proportion of individuals from a defined population group **that can be excluded as a contributor** to an observed DNA mixture

Probability of Exclusion

The probability that a random person (unrelated individual) would be excluded as a contributor to the observed DNA mixture

For each locus, 1 minus the square of the sum of frequencies for the observed alleles

$$PE_i = 1 - \left(\sum_{i=1}^{n} p(A_i) \right)^2$$

Buckleton (2005) *Forensic DNA Evidence Interpretation*, p. 219

Across multiple loci (i.e., **combined probability of exclusion**, CPE):

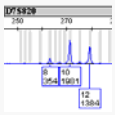
$$PE = 1 - \prod_{l} (1 - PE_l)$$

Buckleton (2005) *Forensic DNA Evidence Interpretation*, p. 221

Combined Probability of Exclusion (CPE)

Each locus is calculated separately and then combined for CPE

$CPE = 1 - (1 - PE_1)(1 - PE_2)(1 - PE_3)\ldots(1\text{-}PE_N)$



Probability of exclusion at a single locus:

- ***The combined frequency of alleles detected (P)***
  P = frequency of allele 1 + frequency of allele 2 + frequency of allele 3, … N

US Caucasian Data

Allele	Frequency
8	0.151
10	0.243
12	0.166

P = 0.151 + 0.243 + 0.166 = **0.56**

- ***The combined frequency of alleles not detected (Q)***
  Q = 1 – P

Q = 1 – 0.56 = **0.44**

- **PE = $Q^2 + 2Q(1-Q)$**

CPI = 1 – CPE

PE = $(0.44)^2 + 2(0.44)(1\text{-}0.44)$ = 0.1936 + 0.4928
PE = 0.686

Calculation from CPI Perspective

Each locus is calculated separately and then combined for CPE

CPI or $P_{profile}$ = (P_{locus1}) (P_{locus2}) (P_{locus3}) … ($P_{locus(N)}$)

Probability of inclusion at a single locus:

- ***Individual frequencies are summed and then squared***
  PI or $P_{locus} = (p_1 + p_2 + p_3 + \ldots + p_N)^2$
  Alleles present in the mixture; Remaining possible alleles in the population

- PE = 1 – P_{locus} = 1 – PI
- **PE = $Q^2 + 2Q(1-Q)$**

Essentially $P^2 + 2PQ + Q^2 = 1$
PI PE

P + Q = 1 so
P = 1 – Q and
Q = 1 – P

Provides probability of an unrelated individual in the population is a contributor to the mixture at the loci examined

Likelihood Ratios

Basic Math Terms

- When '+' is used, this means 'OR'
- When 'x' is used, this means 'AND'
- Pr. is shorthand for probability

- Therefore…
  - the probability of *a* 'AND' *b* happening together is **Pr(*a* and *b*) = *a* x *b***
  - the probability of *a* 'OR' *b* happening together is **Pr(*a* or *b*) = *a* + *b***

Slide information from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Conditioning

- **Probabilities are conditional**, which means that the probability of something is based on a hypothesis

- In math terms, conditioning is denoted by a vertical bar
  - Hence, **Pr(a|b)** means 'the probability of *a* ***given*** that *b* is true"

- The probability of an event *a* is dependent upon various assumptions—and these assumptions or hypotheses can change…

Slide information from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Probability Example – Will It Rain? (1)

Defining the Event and Assumptions/Hypotheses

- Let's suppose that *a* is the probability of an event (e.g., **will it rain?**)
- What is the probability that it will rain in the afternoon – **Pr(a)**?
- This probability is dependent upon assumptions
  - We can look at the window in the morning and observe if it is sunny (s) or cloudy (c)
  - Pr(a) **if** it is sunny (s) is less than Pr(a) **if** it is cloudy (c)
- We can write this as **Pr(a|s)** and **Pr(a|c)**
  - Since sunny or cloudy are the only possibilities, Pr(s) + Pr(c) = 1
  - or **Pr(s) = 1 – Pr(c)**

Slide information from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Probability Example – Will It Rain? (2)

Examining Available Data

- Pr(a|s) and Pr(a|c) can be calculated from data
- How often does it rain in the afternoon when its sunny in the morning?
  - 20 out of 100 observations so **Pr(a|s) = 0.2**
- How often does it rain in the afternoon when it is cloudy in the morning?
  - 80 out of 100 observations so **Pr(a|c) = 0.8**

Slide information from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Probability Example – Will It Rain? (3)

Formation of the Likelihood Ratio (LR)

- The LR compares two probabilities to find out which of the two probabilities is the most likely

The probability that it will rain in the afternoon when it is cloudy in the morning or **Pr(a|c)** is divided by the probability that it will rain in the afternoon when it is sunny in the morning or **Pr(a|s)**

$$LR = \frac{\Pr(a \mid c)}{\Pr(a \mid s)} = \frac{0.8}{0.2} = 4$$

Slide information from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Probability Example – Will It Rain? (4)

Explanation of the Likelihood Ratio

$$LR = \frac{\Pr(a \mid c)}{\Pr(a \mid s)} = \frac{0.8}{0.2} = 4$$

- The probability that it will rain is 4 times more likely **if** it is cloudy in the morning than **if** it is sunny in the morning.
- The word **if** is very important here. It must always be used when explaining a likelihood ratio otherwise the explanation could be misleading.

Slide information from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Likelihood Ratios in Forensic DNA Work

- We evaluate the evidence (*E*) relative to alternative pairs of hypotheses
- Usually these hypotheses are formulated as follows:
  - The probability of the evidence if the crime stain originated with the suspect or Pr(*E*|*S*)
  - The probability of the evidence if the crime stain originated from an unknown, unrelated individual or Pr(*E*|*U*)

$$LR = \frac{\Pr(E \mid S)}{\Pr(E \mid U)}$$

The numerator

The denominator

Slide information from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

The Likelihood Ratio Must Be Stated Carefully

- The probability of the evidence is *x* times more likely **if** the stain came from the suspect Mr. Smith than **if** it came from an unknown, unrelated individual.
- It is not appropriate to say: "The probability that the stain came from Mr. Smith." because we must always include the conditioning statement – i.e., **always make the hypothesis clear in the statement.**
- Always use the word '**if**' when using a likelihood ratio to avoid this trap

Slide information from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Likelihood Ratio (LR)

- Provides ability to express and evaluate both the prosecution hypothesis, H_p (the suspect is the perpetrator) and the defense hypothesis, H_d (an unknown individual with a matching profile is the perpetrator)

$$LR = \frac{H_p}{H_d}$$

- **The numerator, H_p, is usually 1** – since in theory the prosecution would only prosecute the suspect if they are 100% certain he/she is the perpetrator
- The denominator, **H_d**, is typically the profile frequency in a particular population (based on individual allele frequencies and assuming HWE) – i.e., **the random match probability**

Relationship between Likelihood Ratio (LR) and Random Match Probability (RMP)

- For single source samples or deduced major component profiles in a mixture…

$$LR = \frac{1}{RMP} \quad \text{or} \quad RMP = \frac{1}{LR}$$

Example #1

A Single Locus from a 2-Person Mixture

- Consider a simple **two person mixture** with one locus consisting of two heterozygotes with non-overlapping alleles
- If the suspect is *ab*, then there must be another (unknown person) who is *cd*

Suspect = a,b

Forget peak heights for the time being

Adapted from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Example #1

The Two Hypotheses Are Formed…

- **Prosecution (H_p)**: The DNA result has come from the suspect and one unknown person, or **Pr(*E*|*S*,*U*)**
- **Defense (H_d)**: The DNA result has come from two unknown people, or **Pr(*E*|U_1,U_2)**

$$LR = \frac{\Pr(E \mid S,U)}{\Pr(E \mid U_1,U_2)}$$

Suspect = a,b

Forget peak heights for the time being

Adapted from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Example #1

Formulating the Numerator
(Prosecution Hypothesis)

- If the prosecution hypothesis is true, then we would expect genotype *ab* to be present with 100% probability or Pr=1.
- The chance of seeing an unknown person of type *cd* is the frequency of that type in the population or $2p_cp_d$, where p_c is the allele frequency for allele c.
- **Pr(*E*|*S*,*U*)** = 1 x $2p_cp_d$ **= $2p_cp_d$**

Suspect = a,b

Forget peak heights for the time being

Adapted from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Example #1

Formulating the Denominator
(Defense Hypothesis)

- The defense claims that the evidence could come from any two random individuals
- We must work out **all possible pairwise combinations** from alleles *abcd* and their probabilities (genotype frequencies)

Multiplied because you are considering individual #1 **AND** individual #2

Added because you are considering any of the possibilities (combination 1,2,3,4,5, **OR** 6)

Individual #1	Individual #2	Products
ab	cd	$2p_ap_b \times 2p_cp_d$ → $4p_ap_bp_cp_d$
ac	bd	$2p_ap_c \times 2p_bp_d$ → $4p_ap_bp_cp_d$
ad	bc	$2p_ap_d \times 2p_bp_c$ → $4p_ap_bp_cp_d$
cd	ab	$2p_cp_d \times 2p_ap_b$ → $4p_ap_bp_cp_d$
bd	ac	$2p_bp_d \times 2p_ap_c$ → $4p_ap_bp_cp_d$
bc	ad	$2p_bp_c \times 2p_ap_d$ → $4p_ap_bp_cp_d$
Sum of products		**$24p_ap_bp_cp_d$**

Reverse combinations

Pr(*E*|U_1,U_2) = $24p_ap_bp_cp_d$

Adapted from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Example #1

Formulating the Likelihood Ratio

- The numerator and denominator are combined to form the LR
- And common elements in both numerator and denominator are eliminated to simplify the algebraic equation…

$$LR = \frac{\Pr(E \mid S,U)}{\Pr(E \mid U_1,U_2)} = \frac{\cancel{2}\,p_c p_d}{\cancel{24}^{12}\,p_a p_b p_c p_d} = \frac{1}{12\,p_a p_b}$$

All LR Calculations Follow the Same Basic Rules Just Shown

- Form hypotheses
  - Keep in mind what you are conditioning on
- The LR numerator belongs to the prosecution
- The LR denominator belongs to the defense
- Numerator and denominator are combined and equation is simplified
- Allele frequency values are placed into the equation for each locus
- **The LR from each locus is combined through multiplication if the loci are independently inherited (i.e., the product rule) to form a LR for the entire profile**

Example #2

Another Example…

- The evidentiary mixture profile is from a semen stained vaginal swab and possesses alleles a, b, c, and d.
- The suspect is a,b and the victim is c,d.
- Because it is reasonable to assume that the victim's alleles would be present on the swab (i.e., an intimate sample), we can condition on this…

Suspect = a,b
Victim = c,d

Example #2

With an Intimate Sample, the Hypothesis Changes…

- **Prosecution (H_p)**: The DNA result has come from the suspect and the victim, or **Pr(*E*|*S,V*)**
- **Defense (H_d)**: The DNA result has come from the victim and one unknown person, or **Pr(*E*|*U,V*)**

$$LR = \frac{\Pr(E \mid S,V)}{\Pr(E \mid U,V)}$$

Suspect = a,b
Victim = c,d

Example #2

Formulating the Numerator (Prosecution Hypothesis)

- The prosecution hypothesis (S+V) is completely explains the evidence. Hence, the probability is Pr=1
- **Pr(*E*|*S,V*) = 1 x 1 = 1**

Suspect = a,b
Victim = c,d

Example #2

Formulating the Denominator (Defense Hypothesis)

- The defense hypothesis is that the presence of alleles *a* and *b* are the result of an unknown person – and they concede that alleles *c* and *d* come from the victim
- Since the frequency of an unknown, unrelated individual possessing alleles *a* and *b* in the population is $2p_ap_b$, where p_a is the allele frequency for allele *a* and p_b is the allele frequency for allele *b*, then
- **Pr(*E*|*U,V*)** = $2p_ap_b$ x 1 = **$2p_ap_b$**

Suspect = a,b
Victim = c,d

Example #2

Formulating the Likelihood Ratio

- The numerator and denominator are combined to form the LR

$$LR = \frac{\Pr(E \mid S,V)}{\Pr(E \mid U,V)} = \frac{1}{2p_a p_b}$$

- **Note that this LR is the same as for a non-mixed sample comprising the suspect alone.**
- This example then is an illustration of simplification by "subtraction" (victim's alleles are being removed from mathematical consideration…).

Adapted from Peter Gill (ISFG 2007 workshop, Copenhagen, August 20-21, 2007)

Forming the Denominator (H_d) for the LR…

Evidence (Mixture)	Victim	Suspect	LR
A_1, A_2, A_3	A_2, A_3	A_1, A_2	$\frac{1}{p_1(2p_2+2p_3+p_1)}$
8,10,12	10,12	8,10	

Potential Combinations:

If victim is A_2,A_3, then perpetrator could be

	Type	Frequency (probability)
H_d	A_1,A_2	$2p_1p_2$
	A_1,A_3	$2p_1p_3$
	A_1,A_1	p_1^2

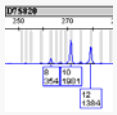
Other possible genotypes contributing to the evidence

Determine joint probability through summing individual probabilities

$2p_1p_2 + 2p_1p_3 + p_1^2 \longrightarrow p_1(2p_2 + 2p_3 + p_1)$

Likelihood Ratio (LR) Calculations

Evidence (Mixture)	Victim	Suspect	LR
A_1, A_2, A_3	A_2, A_3	A_1, A_2	$\frac{1}{p_1(2p_2+2p_3+p_1)}$
8,10,12	10,12	8,10	



US Caucasian Data

Allele	Frequency
A_1 8	p_1 0.151
A_2 10	p_2 0.243
A_3 12	p_3 0.166

$$LR = \frac{1}{(0.151)[(2)(0.243)+2(0.166)+(0.151)]}$$

LR = 6.83

Does not consider peak height information

The prosecution hypothesis (that the suspect is the perpetrator) is 6.83 times more likely than the defense hypothesis (that an unknown, unrelated individual is the perpetrator).

Likelihood Ratios for the Following Hypotheses

H_p: The mixture contains the DNA of the victim and the suspect

H_d: The mixture contains the DNA of the victim and an unknown, unrelated individual

Evidence (Mixture)	Victim	Suspect	LR
A_1, A_2, A_3, A_4	A_1, A_2	A_3, A_4	$\frac{1}{2p_3p_4}$
A_1, A_2, A_3	A_1, A_2	A_1, A_3 or A_2, A_3 or A_3, A_3	$\frac{1}{p_3(2p_1+2p_2+p_3)}$
A_1, A_2, A_3	A_1, A_1	A_2, A_3	$\frac{1}{2p_2p_3}$
A_1, A_2	A_1, A_2	A_1, A_1 or A_1, A_2 or A_2, A_2	$\frac{1}{(p_1+p_2)^2}$
A_1, A_2	A_1, A_1	A_1, A_2 or A_2, A_2	$\frac{1}{p_2(2p_1+p_2)}$
A_1, A_1	A_1, A_1	A_1, A_1	$\frac{1}{p_1^2}$

Adapted from Buckleton (2005) *Forensic DNA Evidence Interpretation*, Table 7.1, p. 229

DAB Recommendations on Statistics

February 23, 2000

Forensic Sci. Comm. 2(3); available on-line at http://www.fbi.gov/hq/lab/fsc/backissu/july2000/dnastat.htm

"The DAB finds either one or both PE or LR calculations acceptable and strongly recommends that one or both calculations be carried out whenever feasible and a mixture is indicated"

- Probability of exclusion (PE)
  - Devlin, B. (1993) Forensic inference from genetic markers. *Statistical Methods in Medical Research*, 2, 241–262.
- Likelihood ratios (LR)
  - Evett, I. W. and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sinauer, Sunderland, Massachusetts.