STATISTICAL INTERPRETATION OF DNA MIXTURES

STATISTICAL INTERPRETATION OF DNA MIXTURES

Guro Dørum

Norwegian University of Life Sciences (UMB)

guro.dorum@umb.no

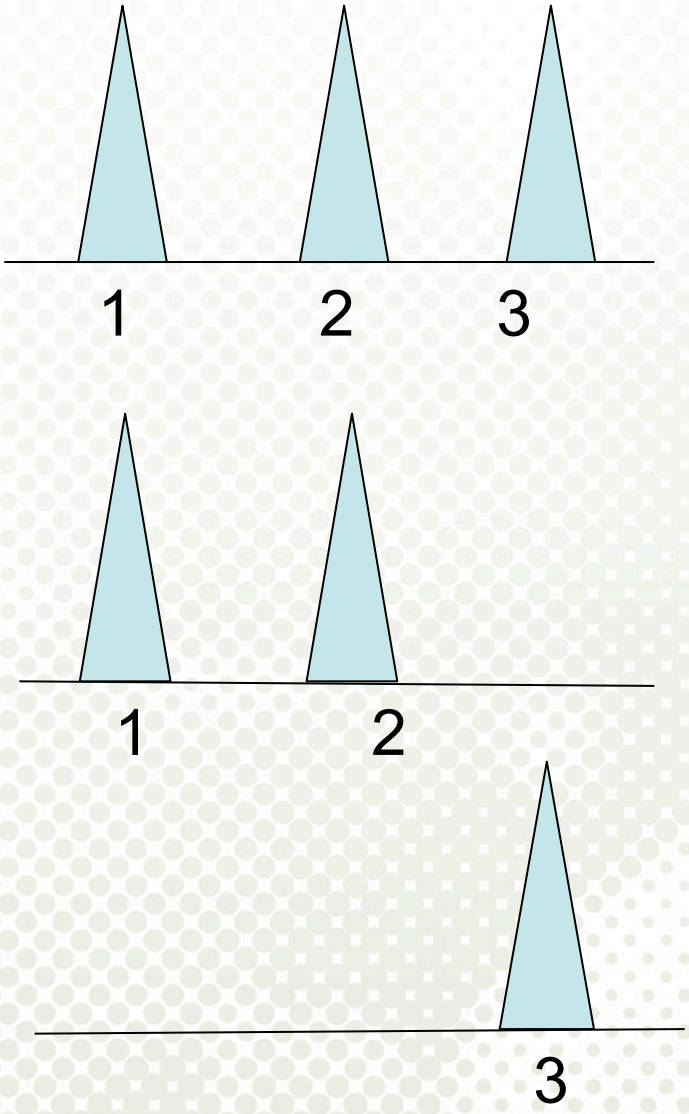
Copenhagen, October 9, 2013

OVERVIEW

- Part I: Basic statistical methods
  - RMNE
  - Binary LR
- Part II: Advanced statistical methods
  - Low template DNA, drop-in and drop-out
  - LR with drop-in and drop-out (semi-continuous LR)
  - Continuous LR

PART I: BASIC STATISTICAL METHODS

MOTIVATING EXAMPLE – MIXTURE, ONE LOCUS



Evidence

Forget about peak heights now

Victim (known contributor)

Suspect

RANDOM MAN NOT EXCLUDED (RMNE)

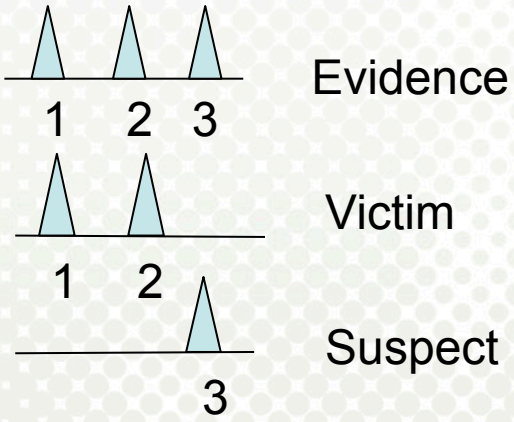
- The fraction of the population not excluded as a contributor / the probability that a random man cannot be excluded as a contributor

With all allele frequencies $p_i = 0.2$

$$RMNE = (p_1 + p_2 + p_3)^2 =$$

$$p_1^2 + p_2^2 + p_3^2 + 2p_1p_2 + 2p_1p_3 + 2p_2p_3$$

$$= 0.36$$



- *Approximately 1 in 3 people could be a contributor to the sample*

LIKELIHOOD RATIO (LR)

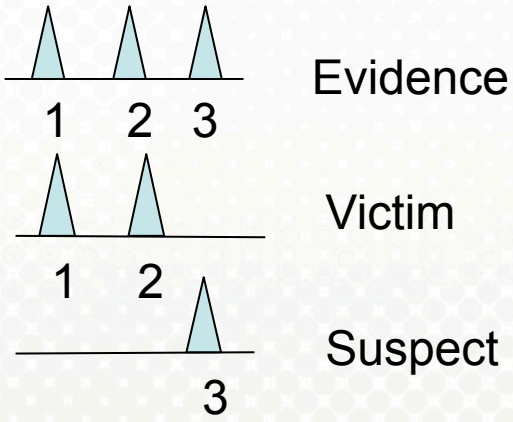
- Ratio of two probabilities
  - The probability of the **evidence (E)** given the prosecution hypothesis (H_p)
  - The probability of the **evidence (E)** given the defense hypothesis (H_d)

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)}$$

- If LR > 1 → The evidence supports H_p
- If LR < 1 → The evidence supports H_d

LIKELIHOOD RATIO

- Specify two hypotheses
  - H_p: Victim + Suspect
  - H_d: Victim + Unknown



- The likelihood ratio, with all allele frequencies p_i=0.2, is

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)} = \frac{1}{p_3^2 + 2p_1p_3 + 2p_2p_3} = \frac{1}{0.2} = 5$$

- The numerator is 1 because a mixture of the victim and suspect gives those evidence alleles with probability 1
- The denominator is the summed probability that the unknown will have one of the genotypes 3,3, 1,3 or 2,3

LIKELIHOOD RATIO

- What does an LR of 5 mean?
- *The evidence is 5 times more likely **IF** the victim and suspect are the contributors than **IF** the victim and an unrelated unknown are the contributors*

- It does not mean that it is 5 times more likely that the victim and the suspect are the contributors than the victim and an unrelated unknown
  - Then we would say something about $P(H_p|E)$ and $P(H_d|E)$, which we do not know!

LIKELIHOOD RATIO

- Baye's theorem

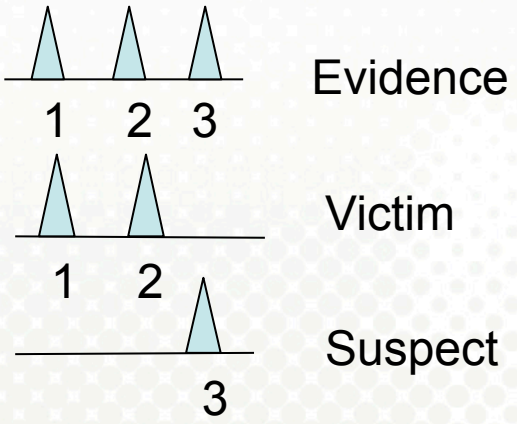
$$\frac{P(H_p)}{P(H_d)} \times \frac{P(E \mid H_p)}{P(E \mid H_d)} = \frac{P(H_p \mid E)}{P(H_d \mid E)}$$

Prior odds — LR — Posterior odds

- To say something about the probability of the hypotheses, we have to make assumptions about their prior probability
- This is only common in paternity cases!

LR DEPENDS ON SUSPECT ALLELE FREQUENCY

- What happens if the suspect's alleles are very rare?
- With frequencies
  p_1=0.3, p_2=0.25, p_3=0.05



$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)} = \frac{1}{p_3^2 + 2p_1p_3 + 2p_2p_3} = \frac{1}{0.032} = 31$$

- In comparison, RMNE is unchanged

$$RMNE = (p_1 + p_2 + p_3)^2 = (0.3 + 0.25 + 0.05)^2 = 0.36$$

- RMNE does not care about reference profiles!

RMNE AND LR – PROS AND CONS

RMNE

+ Simple calculation
+ Easier to explain in court
+ (Requires no assumption of the number of contributors)
÷ **Wastes information**

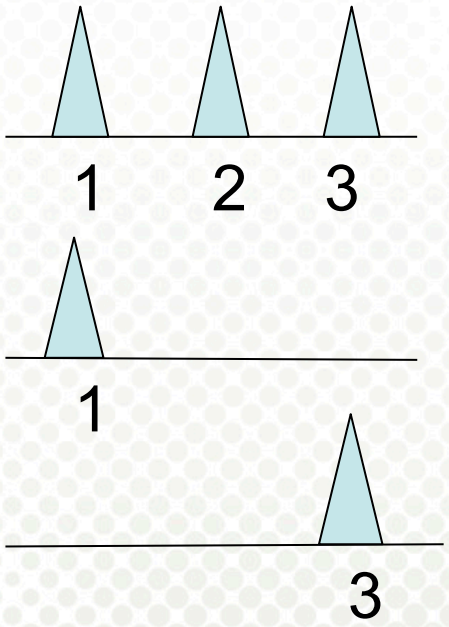
LR

+ Uses all available information
+ Better estimate of weight-of-evidence
+ Phenomena such as drop-out and drop-in can be modelled (which we will soon look at)
÷ **More difficult to explain in court**

Likelihood ratios (LR) are recommended for mixtures by ISFG DNA Commission

See also: *A discussion of the merits of random man not excluded and likelihood ratios*, J. Buckleton and J. Curran (2008) FSI Genetics

RMNE MAKES NO ASSUMPTION ABOUT THE NUMBER OF CONTRIBUTORS



Evidence

Victim (known contributor)

Suspect

- If we assumed two contributors, suspect would have been excluded
- However RMNE does not care:

$$RMNE = (p_1 + p_2 + p_3)^2 = p_1^2 + p_2^2 + p_3^2 + 2p_1p_2 + 2p_1p_3 + 2p_2p_3$$

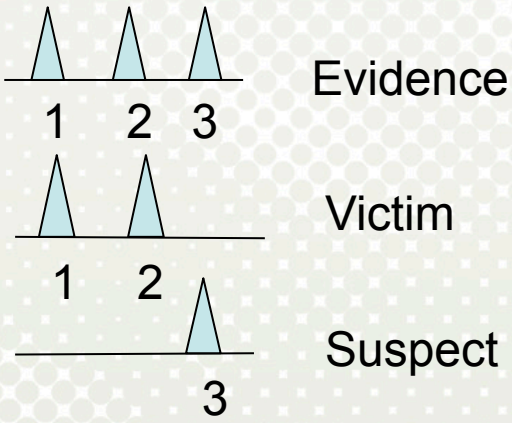
RANDOM MATCH PROBABILITY (RMP)

- The probability that a random man cannot be excluded as a contributor, conditioned on the reference profile(s) and assumed number of contributors

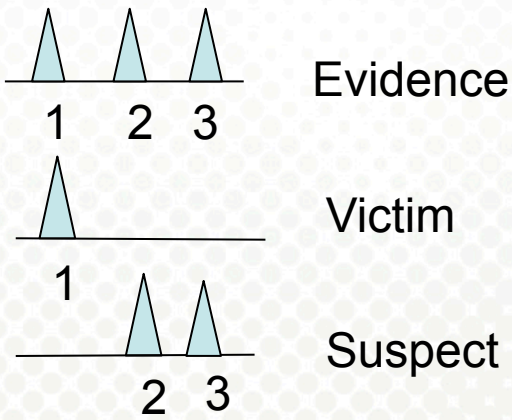
- Assuming two contributors, the victim explains 1 and 2 so other contributor must explain 3

- RMP = $p_3^2+2p_1p_3+2p_2p_3 = 0.2$

- Note that: LR = 1/RMP



EXERCISE: CALCULATE RMNE AND LR



1) Allele frequencies $p_1=0.2$, $p_2=0.1$, $p_3=0.05$
2) Allele frequencies $p_1=0.05$, $p_2=0.1$, $p_3=0.2$

- Calculate RMNE and LR for the hypotheses H_p: Victim + suspect vs. H_d: Victim + Unknown

SOLUTION TO EXERCISE

1) Allele frequencies p_1=0.2, p_2=0.1, p_3=0.05

$$RMNE = (p_2 + p_2 + p_3)^2 = (0.2 + 0.1 + 0.05)^2 = 0.1225$$

$$LR = \frac{1}{2p_2p_3} = \frac{1}{2 \times 0.1 \times 0.05} = 100$$

2) Allele frequencies p_1=0.05, p_2=0.1, p_3=0.2

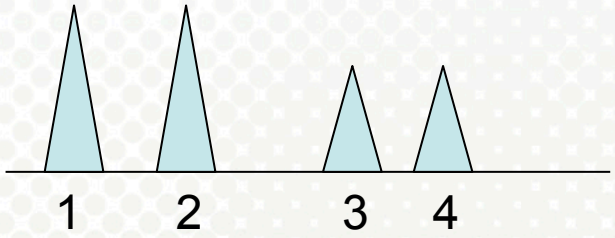
$$RMNE = (p_2 + p_2 + p_3)^2 = (0.05 + 0.1 + 0.2)^2 = 0.1225$$

$$LR = \frac{1}{2p_2p_3} = \frac{1}{2 \times 0.1 \times 0.2} = 25$$

PART II: ADVANCED STATISTICAL METHODS

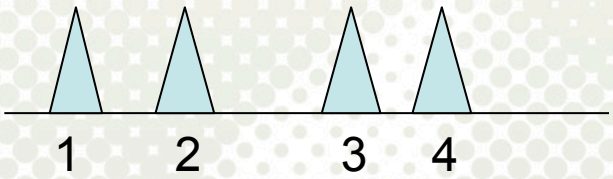
HIGH AND LOW TEMPLATE DNA

- High template DNA
  - The epg reflects the composition of the sample



Possible genotypes
1,2 3,4

- Low template DNA
  - The epg does not reflect the composition of the sample that well



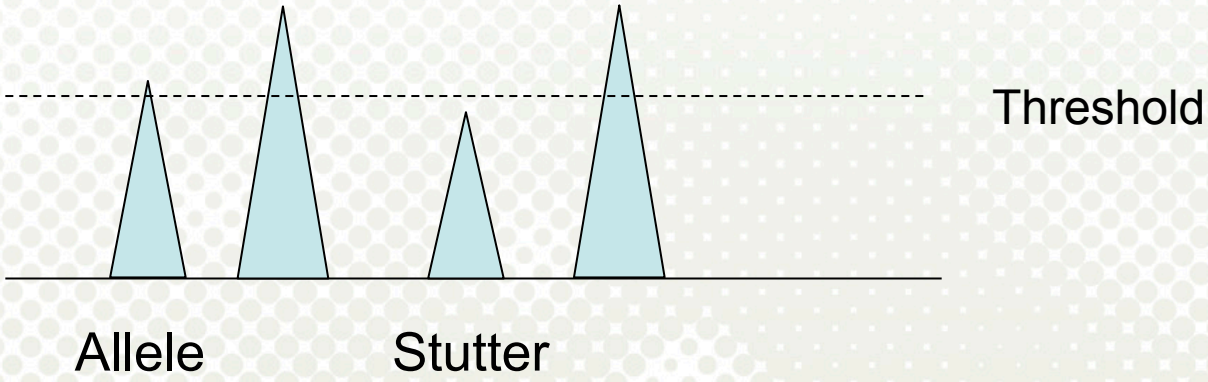
Possible genotypes

1,2 3,4	3,4 1,2
1,3 2,4	2,4 1,3
1,4 2,3	2,3 1,4

LOW TEMPLATE DNA

- Heterozygous imbalance

- Stutter or allele?



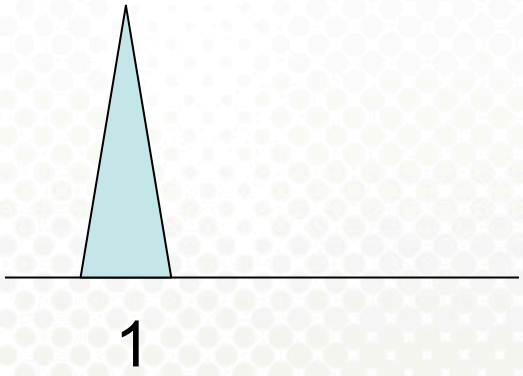
- Much less information in the peak areas

DROP-IN AND DROPOUT

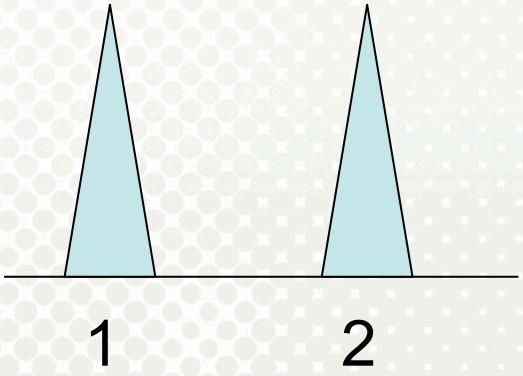
- Drop-in and dropout are stochastic phenomena that appears in low template DNA
- Dropout: an allele fails to amplify (below detection level)
  - An allele present in the reference profile but not in the evidence
- Drop-in: one or two foreign alleles in the profile
  - An allele present in the evidence but not in the reference profile
- Drop-in is not the same as contamination
  - Gross contamination can be dealt with by including an additional unknown contributor
  - Drop-in events are considered independent while contamination is dependent

DROP-OUT

- Evidence



- Suspect

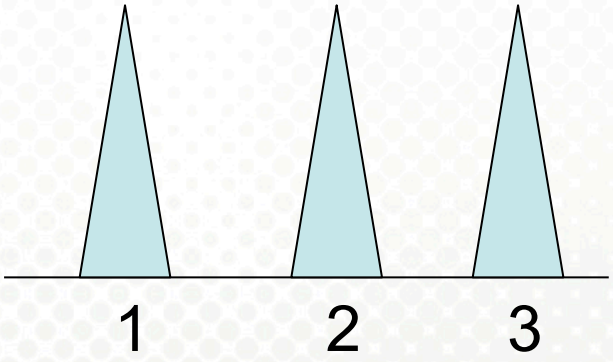


Match? Then allele 2 must have dropped out of the evidence

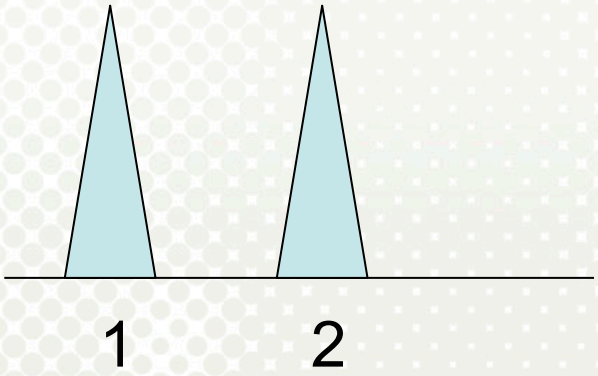
Gill ISFG 2013 workshop

DROP-IN

- Evidence



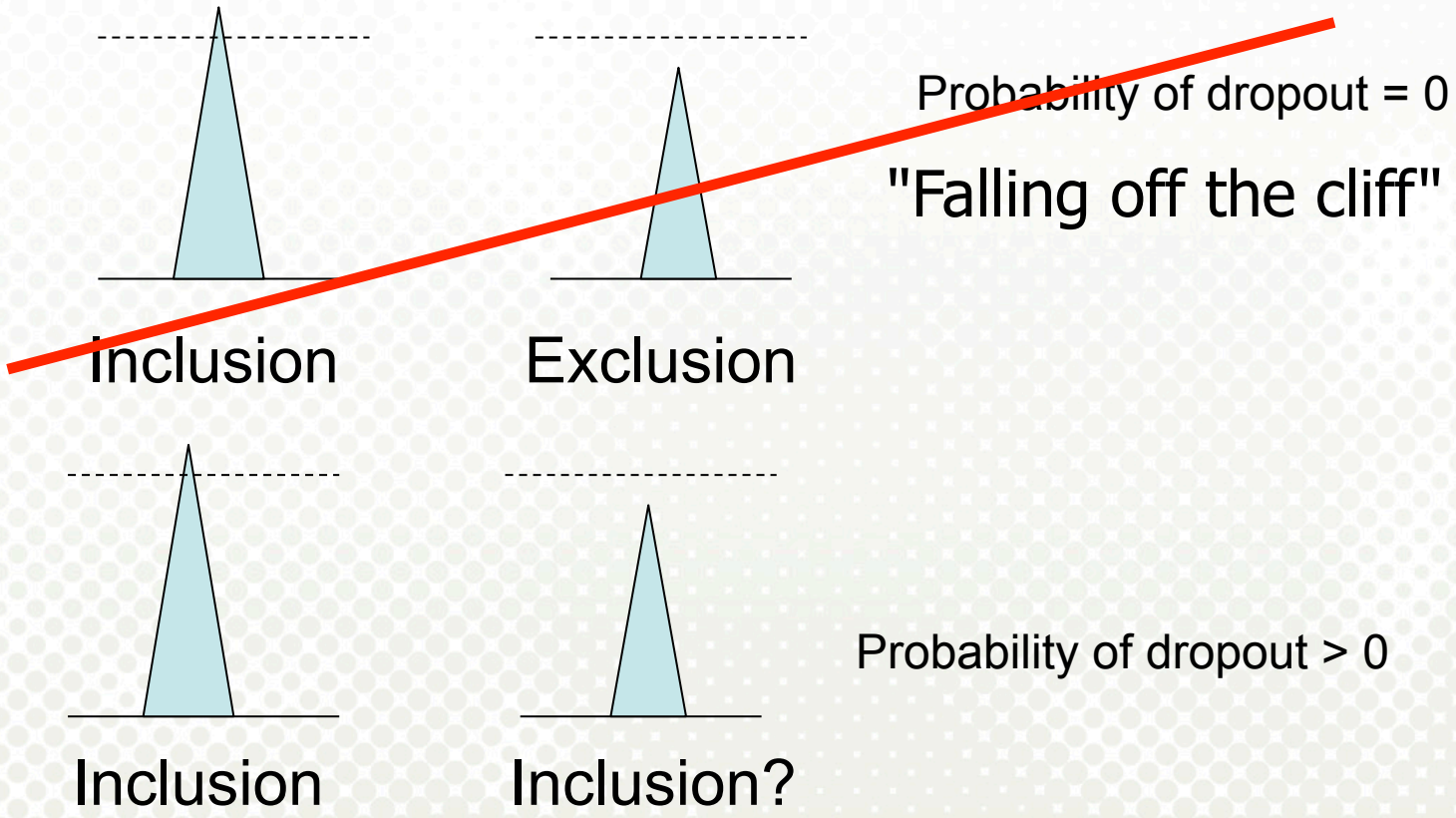
- Suspect



Match?
Then allele 3 must have dropped into the evidence

Gill ISFG 2013 workshop

THE EFFECT OF ACCOUNTING FOR DROP-IN AND DROPOUT



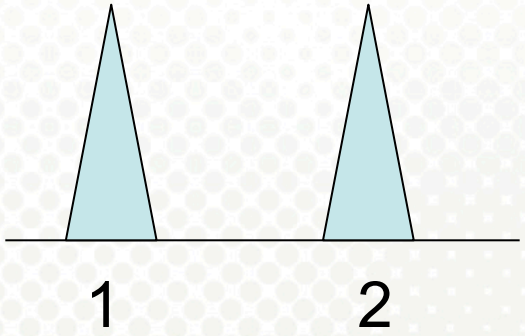
- Introduce uncertainty about whether the allele is present or not
- No need to define a profile as included or excluded!

We need models that can incorporate uncertainties about the data

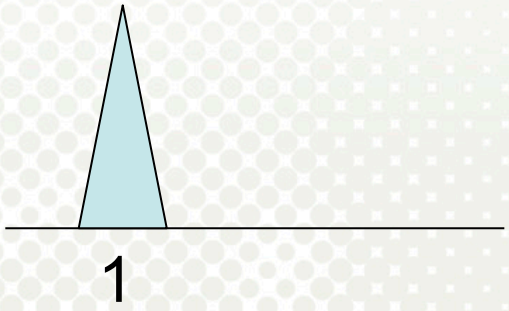
Here we will focus on LR with drop-in and dropout (semi-continuous LR)

EXAMPLE 1: DROP-OUT

- Single-source for simplicity, but equally applicable to mixtures
- Suspect



- Evidence



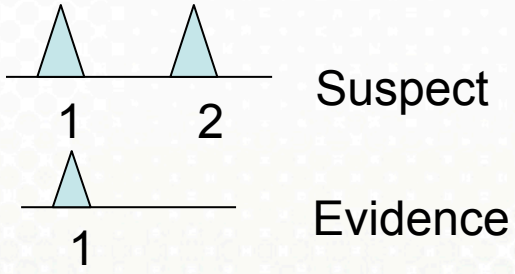
$$\text{Binary LR} = \frac{0}{p_1^2}$$

$$\text{Semi-continuous LR} = \frac{?}{p_1^2}$$

H_p: Suspect
H_d: Unknown

? = some number between 0 and 1

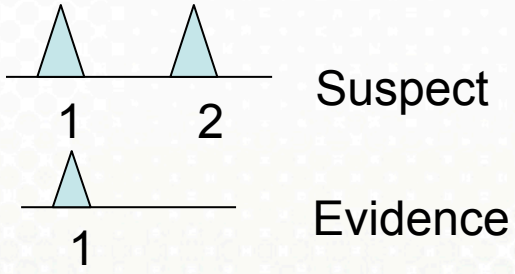
EXAMPLE 1: LR WITH DROPOUT



- H_p: Suspect
- H_d: Unknown

- We define
  - probability of dropout for heterozygotes = *d*
    - *both alleles drop out with probability* d^2
  - probability of dropout for homozygotes = *d′*
  - where $d' \leq d^2$

EXAMPLE 1: LR WITH DROPOUT



- H_p: Suspect
- H_d: Unknown

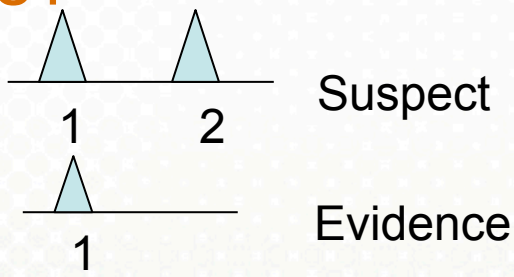
- For H_p to be true
  - allele 2 in the suspect's profile has dropped out
  - allele 1 in the suspect's profile has not dropped out

- The probability of the evidence given H_p is

$$P(E \mid H_p) = P(\text{dropout of } 2) \times \text{P(non-dropout of } 1)$$
$$= d \times (1 - d)$$

EXAMPLE 1: LR WITH DROPOUT

- H_p: Suspect
- H_d: Unknown



- For H_d to be true the unknown must have a genotype that contains the allele 1
- Let Q denote any other allele than the allele in the evidence
- Assuming that the locus has five alleles, each with frequency p_i

$Q = \{2,3,4,5\}$ and $p_Q = p_2+p_3+p_4+p_5$

EXAMPLE 1: LR WITH DROPOUT

- The genotypes the unknown may have are

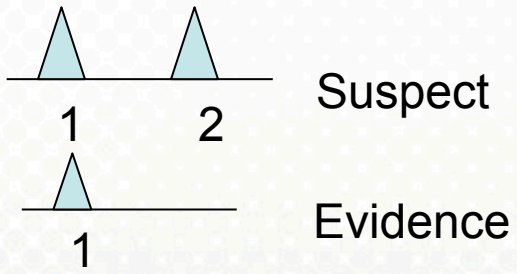
Genotype	Dropout	Genotype probability
1,1	(1-d')	p_1^2
1,Q	(1-d)d	$2p_1p_Q$

- The genotype probability, together with the probability for dropout required to explain the genotype being in the evidence, is the genotype's "weight"
- The probability of the evidence given H_d is

$$P(E \mid H_d) = p_1^2(1-d') + 2p_1p_Q(1-d)d$$

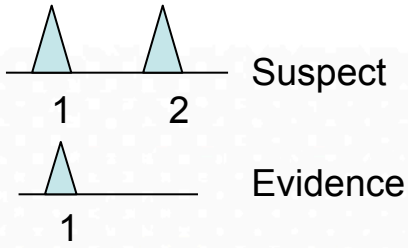
EXAMPLE 1: LR WITH DROPOUT

- H_p: Suspect
- H_d: Unknown



- Putting it all together

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)} = \frac{d(1-d)}{p_1^2(1-d') + 2p_1p_Q d(1-d)}$$



EXAMPLE 1: LR WITH DROPOUT

- What is the effect of dropout probability and allele frequency on LR?
- Scenario 1: p_1=0.2, p_Q=0.8, d=0.05, d'=0.05^2

$$LR = \frac{0.05(1-0.05)}{0.2^2(1-0.05^2)+2\times 0.2\times 0.8\times 0.05(1-0.05)} = 0.86$$

- Scenario 2: p_1=0.2, p_Q=0.8, d=0.3 and d'=0.3^2

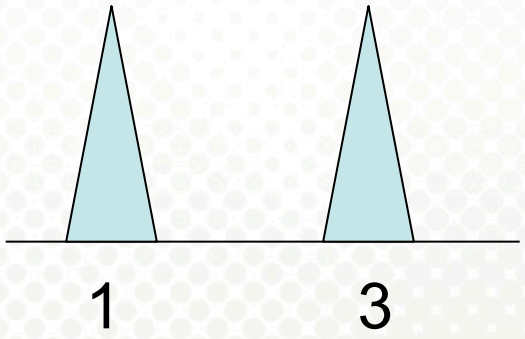
$$LR = \frac{0.3(1-0.3)}{0.2^2(1-0.3^2)+2\times 0.2\times 0.8\times 0.3(1-0.3)} = 2.03$$

- Scenario 3: p_1=0.01, p_Q=0.99, d=0.05, d'=0.05^2

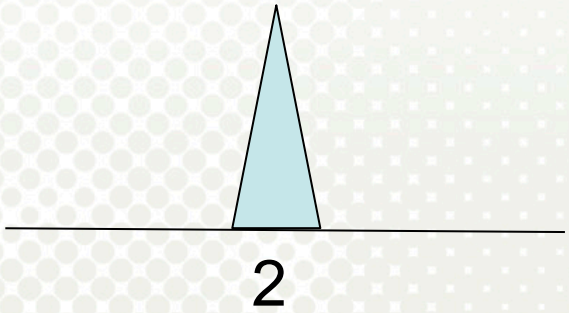
$$LR = \frac{0.05(1-0.05)}{0.01^2(1-0.05^2)+2\times 0.01\times 0.99\times 0.05(1-0.05)} = 45.66$$

EXAMPLE 2: DROPOUT AND DROP-IN

- Suspect



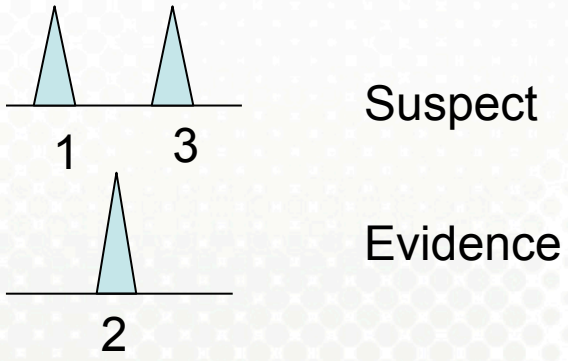
- Evidence



H_p: Suspect
H_d: Unknown

EXAMPLE 2: LR WITH DROP-IN AND DROPOUT

- H_p: Suspect
- H_d: Unknown



- For H_p to be true
  - both alleles 1 and 3 in the suspect's profile must have dropped out
  - allele 2 in the evidence must have dropped in
- We define the probability of drop-in = *c*. The probability of drop-in of allele 2 is cp_2
- The probability of the evidence given H_p is

$$P(E \mid H_p) = P(\text{dropout of } 1) \times \text{P}(\text{dropout of } 3) \times \text{P}(\text{drop-in of } 2)$$
$$= d \times d \times cp_2$$

EXAMPLE 2: LR WITH DROP-IN AND DROPOUT

- Under H_d, the unknown may have any genotype

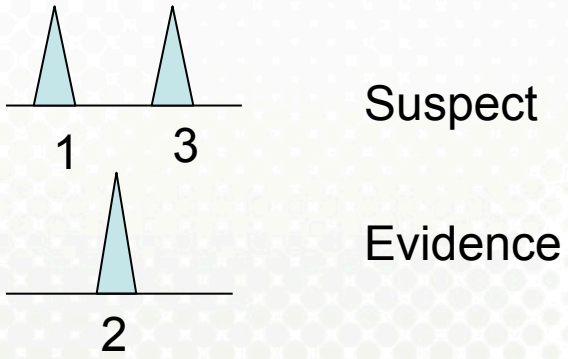
Genotype	Dropout	Drop-in	Genotype probability
2,2	(1-d')	(1-c)	p_2^2
2,Q	(1-d)d	(1-c)	$2p_2p_Q$
Q,Q	d'	cp_2	p_Q^2

- The probability of the evidence given H_d is

$$P(E \mid H_d) = p_2^2(1-d')(1-c) + 2p_2p_Q(1-d)d(1-c) + p_Q^2 d' cp_2$$

EXAMPLE 2: LR WITH DROP-IN AND DROPOUT

- H_p: Suspect
- H_d: Unknown

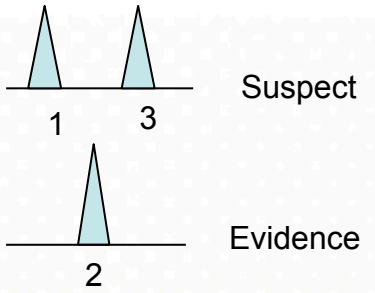


- Putting it all together

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)}$$

$$= \frac{d^2 c p_2}{p_2^2(1-d')(1-c) + 2p_2 p_Q (1-d) d (1-c) + p_Q^2 d' c p_2}$$

EXAMPLE 2: LR WITH DROP-IN AND DROPOUT



- What is the effect of drop-in, dropout and allele frequency?
- Scenario 1: p_2=0.1, p_Q=0.9, d=0.3, d'=0.3^2, c=0.05

$$LR = \frac{0.3^2 \times 0.05 \times 0.1}{0.1^2(1-0.3^2)(1-0.05) + 2 \times 0.1 \times 0.9(1-0.3)0.3(1-0.05) + 0.9^2 \times 0.05 \times 0.1} = 0.01$$

- Scenario 2: p_2=0.5, p_Q=0.5, d=0.3, d'=0.3^2, c=0.05

$$LR = \frac{0.3^2 \times 0.05 \times 0.5}{0.5^2(1-0.3^2)(1-0.05) + 2 \times 0.5 \times 0.5(1-0.3)0.3(1-0.05) + 0.5^2 \times 0.05 \times 0.5} = 0.007$$

- Scenario 3: p_2=0.5, p_Q=0.5, d=0.8, d'=0.8^2, c=0.5

$$LR = \frac{0.8^2 \times 0.5 \times 0.5}{0.5^2(1-0.8^2)(1-0.5) + 2 \times 0.5 \times 0.5(1-0.8)0.8(1-0.5) + 0.5^2 \times 0.5 \times 0.5} = 1.28$$

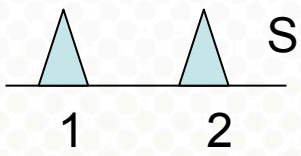
SOFTWARE NEEDED!

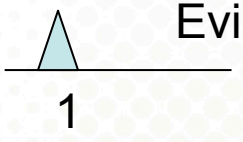
- The two examples considered here are very simplified
- The LR calculations quickly gets complicated with both drop-in and drop-out accounted for
- That is why we have software (like LRmix) to do these calculations!

LR WITH DROP-IN AND DROPOUT

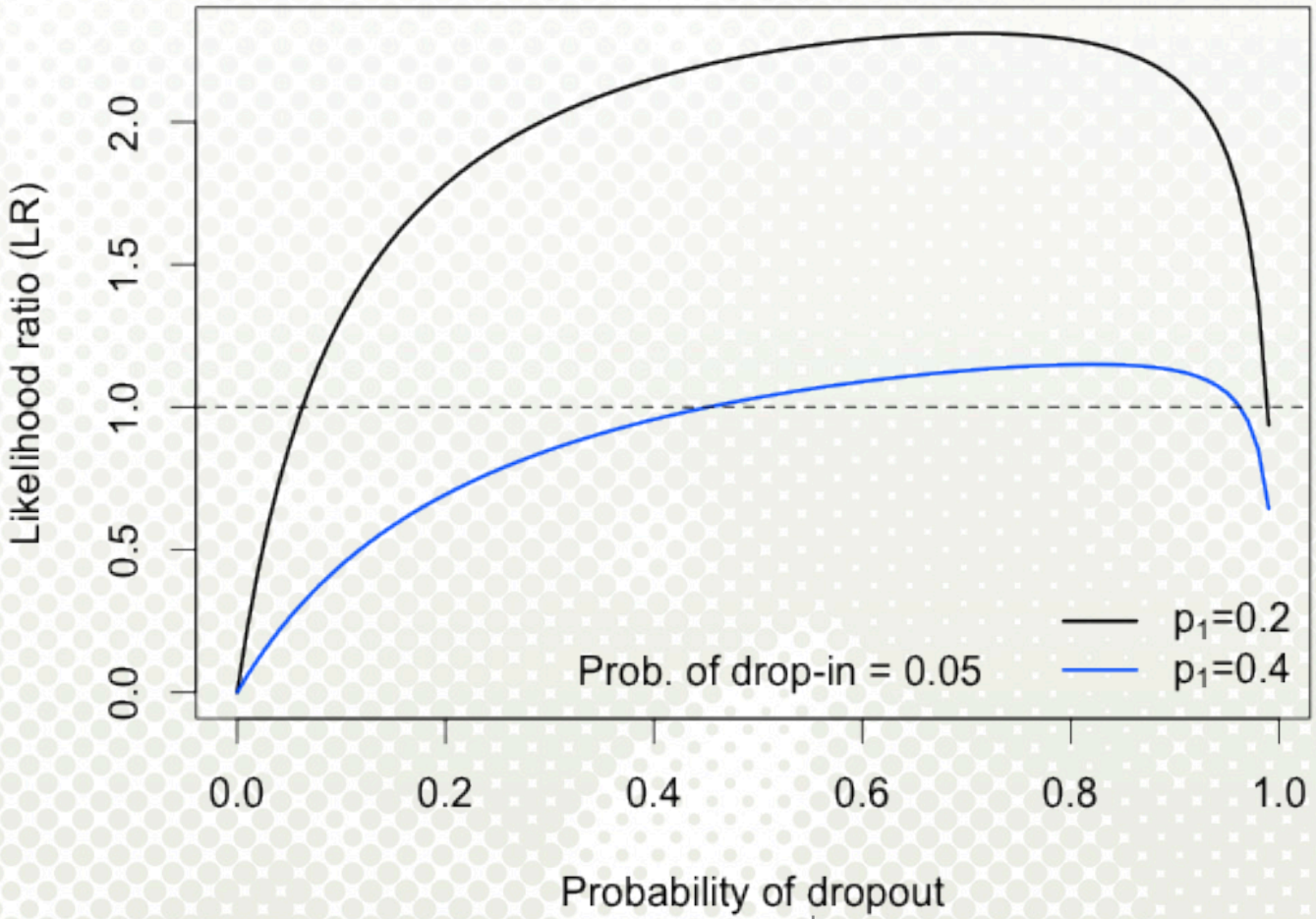
- This LR model was first described by Curran et al. (2005)
- It is the basis for the LRmix software
- We may think of this LR as "semi-continuous"
  - Information from epg is included in LR as uncertainties in the data through drop-in and dropout probabilities
- Other software include likeLTD (David Balding) and FST (NYOCME, Mitchell *et al.*)

THE EFFECT OF DROPOUT ON LR





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



RECOMMENDATIONS OF THE ISFG DNA COMMISSION

13. Summary of recommendations of the ISFG DNA commission

(1) Probabilistic methods following the '*basic model*' described here can be used to evaluate the evidential weight of DNA results considering drop-out and/or drop-in.
(2) Estimates of drop-out and drop-in probabilities should be based on validation studies that are representative of the method used.
(3) The weight of the evidence should be expressed following likelihood ratio principles.
(4) The use of appropriate software is highly recommended to avoid hand-calculation errors.

Gill et al. (2012) DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may included drop-out and/or drop-in using probabilistic methods. FSI Genetics

RECOMMENDATIONS OF THE ISFG DNA COMMISSION

- "(…) it should be noted that the approach described here still requires a rigid assessment of the overall quality of a given DNA profile and its suitability for further analysis based on criteria described in the laboratory's quality management guidelines."

WHAT ABOUT PEAK HEIGHTS?

- Continuous LR methods try to model heterozygous balance, stutters, mixture ratio
- Software include TrueAllele (Mark Perlin, Cybergenetics) and STRmix (ESR (NZ) and Australian collaboration)
- Uses more information → more computationally challenging and time consuming
- "Black box"
  - Not possible to compare with calculations done by hand

SUMMARY - STATISTICAL METHODS

- Basic statistical methods
  - RMNE
  - RMP
  - Binary LR

Binary
"Exclusion or inclusion"

- Advanced statistical methods
  - Semi-continuous LR
  - Continuous LR

Probabilistic
"Everything is possible"

EXERCISE: CALCULATE LR IN EXAMPLE 1 AND 2

1) Example 1 with dropout
   - Probability of dropout for heterozygotes d=0.1
   - Probability of dropout for homozygotes d'=0.1^2
   - Allele frequencies $p_1=p_2=p_3=p_4=p_5=0.2$
2) Example 2 with drop-in and dropout
   - Probability of drop-in c=0.05
   - Same dropout probabilities and allele frequencies as example 1

- Calculate by hand and compare with LRmix

SOLUTION LR EXAMPLE 1

- Define $p_Q = p_2 + p_3 + p_4 + p_5 = 4 \times 0.2 = 0.8$

$$P(E \mid H_p) = 0.1 \times (1 - 0.1) = 0.09$$

$$P(E \mid H_d) = 0.2^2(1 - 0.1^2) + 2 \times 0.2 \times 0.8(1 - 0.1)0.1 = 0.0684$$

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)} = \frac{0.09}{0.0684} = 1.3158$$

SOLUTION LR EXAMPLE 2

- Define $p_Q = p_1 + p_3 + p_4 + p_5 = 4 \times 0.2 = 0.8$

$$P(E \mid H_p) = 0.1 \times 0.1 \times 0.05 \times 0.2 = 0.0001$$

$$\begin{aligned} P(E \mid H_d) &= 0.2^2(1-0.1^2)(1-0.05) \\ &+2 \times 0.2 \times 0.8(1-0.1)0.1(1-0.05) \\ &+0.8^2 \times 0.1^2 \times 0.05 \times 0.2 \\ &= 0.0650 \end{aligned}$$

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)} = \frac{0.0001}{0.0650} = 0.0015$$

THE LRMIX PERFORMANCE TEST

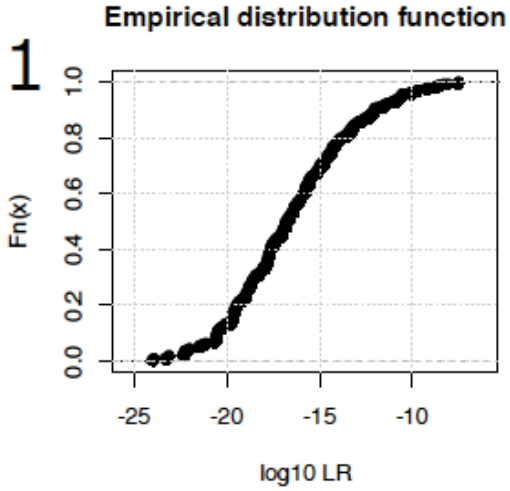
PERFORMANCE TEST / NON-CONTRIBUTOR TEST

- LRmix does a performance test to assess the "robustness" of the LR
- What happens if we replace the suspect with a random man?
- Example with two suspects evaluated in the same hypothesis
  - H_p: V + S1 + S2 vs. H_d: V + Two unknowns

  → log10(LR) = 5.66
- Replace S1 with e.g. 1000 random men drawn from allele frequency database and recalculate LR
- Repeat procedure with S2

PERFORMANCE TEST / NON-CONTRIBUTOR TEST

S1

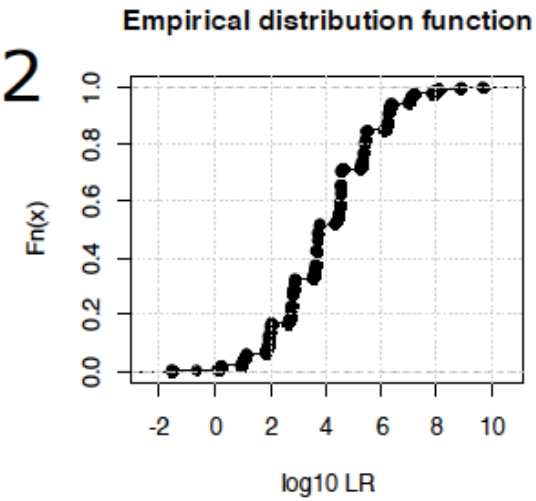


quantile" "value"
"min" "-24.0269"
"0.01" "-23.2479"
"0.05" "-21.4325"
"0.5" "-16.7792"
"0.95" "-10.5699"
"0.99" "-8.4826"
"max" "-7.4584"

b

Original
LR=5.66

S2



"quantile" "value"
"min" "-1.591"
"0.01" "0.126"
"0.05" "1.0629"
"0.5" "3.7167"
"0.95" "7.0392"
"0.99" "7.9833"
"max" "9.6998"

PERFORMANCE TEST / NON-CONTRIBUTOR TEST

- Summarise the plots with the one percentile, median and 99 percentile of the distribution
  - For S1 these are (-23,-16,-8)
  - For S2 these are (0.1,3.7,7.9)
- Remember that calculated log10(LR) = 5.66
- S2 contributes no more to LR than a random man would!
  - No evidence for S2 under H_p
- Be careful with complex hypotheses – the likelihood ratio does not reflect the relative contributions of S1 and S2