

# DNA Fingerprinting Probabilities

---

## I. The Jury's Problem

### A. Decide between **alternatives**

Is the suspect ( $\mathcal{S}$ ) Guilty or Innocent? But the choice must be made when there are many contingencies (motive, alibi, unreliable witnesses, etc).

B. These features suggest that the jury must evaluate *conditional probabilities*: What is the probability of guilt, **given** that certain events occurred?

## II. Conditional Probabilities in Forensics

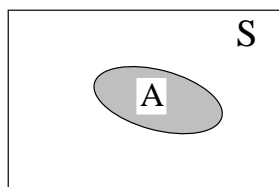
### A. Basic Probability

1. If  $\mathbf{A}$  is an event (e.g., 3 heads in a row from a fair coin), what is  $\mathbf{P}(\mathbf{A})$  (the probability of  $\mathbf{A}$ )?

Let  $\mathbf{S}$  be the **sample space**: the set of all possible outcomes associated with the event. The size of  $\mathbf{S}$  is  $|\mathbf{S}|$ . Let the size of  $\mathbf{A}$  be  $|\mathbf{A}|$ . Then the probability of  $\mathbf{A}$  occurring is:

$$P(A) = \frac{\text{size of } A}{\text{size of } S} = \frac{|A|}{|S|}$$

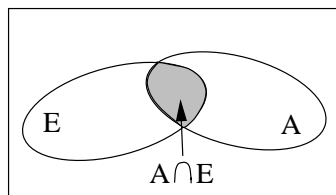
Graphically, this is:



2. Let  $\mathbf{E}$  be another event of interest that has the same sample space as before. Then

$$P(E) = \frac{|E|}{|S|}$$

3. Suppose  $\mathbf{A}$  can occur if  $\mathbf{E}$  has also occurred. What is the probability of  $\mathbf{A}$  occurring given that  $\mathbf{E}$  has already occurred? This is the conditional probability of  $\mathbf{A}$ :  $P(A|E)$ . " $\mathbf{A}$  given  $\mathbf{E}$ " implies that  $\mathbf{A}$  and  $\mathbf{E}$  occur, i.e., the intersection of  $\mathbf{A}$  and  $\mathbf{E} = A \cap E$ . Graphically, the size of the conditional probability is this intersection:



For example, suppose we have two coins: What is the probability of getting a H on one coin given that the other coin is a T?

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

$\mathbf{E} = \text{T in second coin}$

$\mathbf{A} = \text{H in first coin}$

$(A \cap E) = (H, T)$ , but to get the probability we must divide by the set of possibilities:  $|E| = 2 = (H, T)$  and  $(T, T)$ . So,

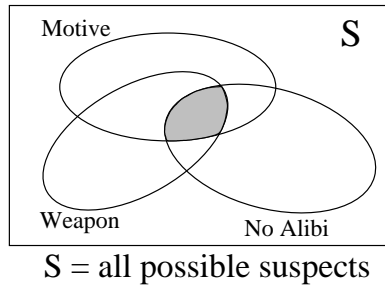
$$P(A|E) = \frac{|A \cap E|}{|E|} = \frac{1}{2}$$

# DNA Fingerprinting Probabilities

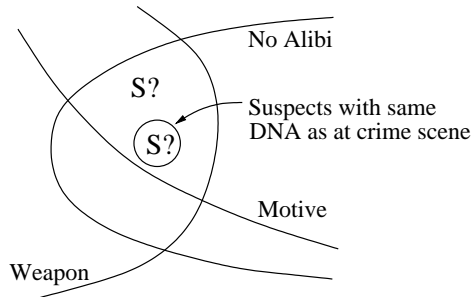
---

## B. Conditional probability applied to crime forensics

1. What is the probability that the suspect is guilty?
2. What is the sample space?
3. What are the contingent events that determine the set of possibly guilty people?
4. What do the intersections mean?



## C. Where does DNA play a role? Narrows set of suspects immensely.



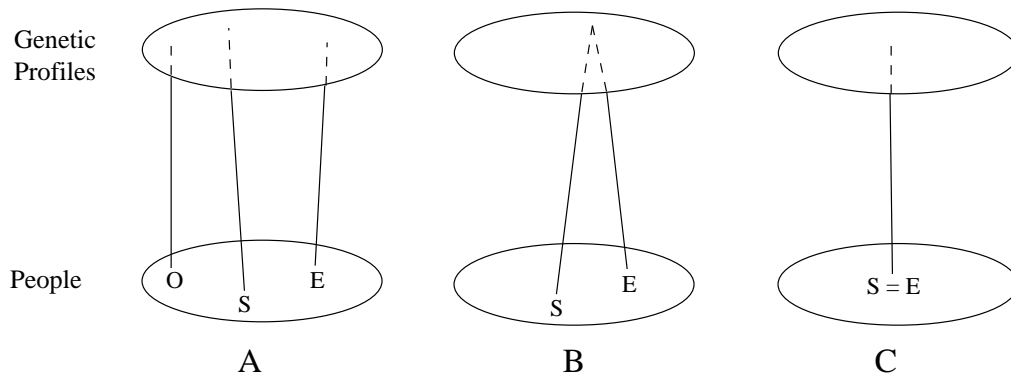
Where does the suspect ( $S$ ) lie?

## D. Jury would like to know:

1. What is the probability that  $S$  is in the circle?
2. How big is the circle?

## III. Computing the Probabilities

- A. Focus on DNA: prosecutor has established motive, access to weapon, no alibi, etc.
- B. What are the possible relationships between suspects and DNA samples?



Bottom ellipses are all possible people that the prosecutor has established as possible suspects. Top ellipses are the genetic profiles (e.g., gels) of all of the suspects.  $S$  is the person being tried;  $E$  is the person that did the crime. We can know the top ellipse

## DNA Fingerprinting Probabilities

---

(evidence), but the jury must decide if the suspect was the person that did the crime (bottom ellipse).

**A** means the suspect was not at the crime scene (no trial!)

**B** means the suspect was not at the crime scene, but **DNA says YES**

**C** means the suspect was at the crime scene, and **DNA says yes**

### C. Notation

$\mathcal{S}$  = suspect                       $\mathcal{E}$  = person at scene leaving evidence  
 $G_s$  = DNA profile of  $\mathcal{S}$        $G_e$  = DNA profile of  $\mathcal{E}$

### D. The possibilities

1.  $G_s \neq G_e \Rightarrow \mathcal{S}$  not at scene
2.  $H_0$ :  $G_s = G_e$  and  $\mathcal{S} \neq \mathcal{E}$
3.  $H_1$ :  $G_s = G_e$  and  $\mathcal{S} = \mathcal{E}$

### E. Terminology for $H_0$ different than inferential statistics.

### F. Jury wants to know:

*What is the probability that  $H_1$  is true relative to  $H_0$ ?*

1. Not the same thing as inferential statistics: In inferential statistics: *What is the probability that two samples have the observed difference in means given that they were drawn from the same population distribution?*  
This is the null hypothesis ( $H_0$ ).
2. Inferential statistics does not allow us to say anything about the alternatives.

## IV. Computing What the Jury Wants.

### A. Which of the two alternatives ( $H_0$ and $H_1$ ) is more likely?

How can we objectively quantify the likelihood of one alternative relative to another?

### B. Example

What is the likelihood that a particular die is fair or not fair? (Fair = all faces equally probable).

Suppose we perform an experiment: Roll a die 5 times and observe two (2) three's. This is a Bernoulli trial. The probability of observing the experimental example follows a **Binomial** probability distribution, which is:

$$b(x; n, \theta) = \text{BUFYCLU: Big Ugly Formula You Can Look Up}$$

$b$ for Binomial	$x$ : # of occurrences of a face in the trial	$n$ : # of rolls of die	$\theta$ : probability that the face will appear
---------------------	---	-------------------------------	--

Our sample:  $x = 2$ ,  $n = 5$ ,  $\theta = ?$ : unknown but may be  $1/6=0.1667$ .

### C. Questions for the die

1. How likely is this **particular** sample?
2. How does this likelihood compare to the likelihood of an unfair die?

## DNA Fingerprinting Probabilities

---

### D. Definition

*Likelihood (of a sample): the probability that a particular sample would be drawn from a specified distribution with particular parameters*

$\mathcal{L}$  = the likelihood

### E. Our Case

Question 1: What is  $\mathcal{L}(\text{two 3's}|\text{binomial with } n = 5, \theta = 0.1667)$  ?

Question 2: What is  $\mathcal{L}(\text{two 3's}|\text{binomial with } n = 5, \theta = 0.1, \theta = 0.5, \text{etc.})$  ?

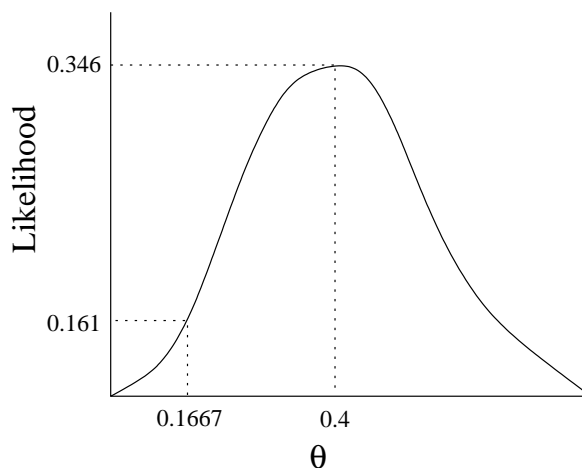
To answer these questions, we need to know how  $\mathcal{L}$  depends on  $\theta$ . We will, therefore, define a **likelihood function** in which  $\theta$  is the independent variable. The function will be dependent on our particular experimental results so the likelihood function for our die experiment is:

$\mathcal{L}(\text{two 3's}|\text{binomial with } n = 5, \theta = \text{variable})$

### F. The likelihood function for the die is:

$$\mathcal{L}(\text{two 3's}|b(2; 5, \theta)).$$

When you substitute into the equation of the binomial distribution (BUFYCLU) different values of  $\theta$ , the plot of this function against  $\theta$  (the unknown probability of a face showing on top) is:



As the plot shows, the most likely value of  $\theta$  is 0.4. So, is the die fair? We must choose between two hypotheses: (1) the die is fair and  $\theta = 0.1667$  and (2) the die is biased with  $\theta = 0.4$ . We'll answer this question later, but for the moment, an important index of the differences between two competing hypotheses is the ratio of their two likelihood values. This is called the likelihood ratio ( $\mathcal{LR}$ ).

### V. $\mathcal{LR}$ Applied to DNA Fingerprinting

#### A. Remember: 2 competing hypotheses:

$H_0: G_s = G_e$  and  $\mathcal{S} \neq \mathcal{E}$ .

The DNA of the suspect is the same as the evidence left at the crime scene and the suspect is not the person leaving the evidence.

$H_1: G_s = G_e$  and  $\mathcal{S} = \mathcal{E}$ .

The DNA of the suspect is the same as the evidence left at the crime scene and the suspect is the person leaving the evidence.

## DNA Fingerprinting Probabilities

---

- B. Use  $\mathcal{LR}$ , which means we must calculate the  $\mathcal{L}$  of  $H_0$  and of  $H_1$  and then calculate the ratio :  $\mathcal{LR} = \mathcal{L}(H_1)/\mathcal{L}(H_0)$ .
- C. Since we are only interested in the case  $G_s = G_e$ , there is only **one** type of DNA, which we will call this DNA  $G$ .
- D. First, calculate  $\mathcal{L}$  of  $H_1$ .

$\mathcal{L}(H_1) =$  what is the probability that the suspect will have DNA =  $G$ , **given** that the evidence DNA= $G$  and (hypothesis)  $\mathcal{S} = \mathcal{E}$  (the suspect did actually leave the evidence).

Intuitively, the probability must 1.0. (The question is: what is the probability that X is true, given that X is true?). But we can check this truism using conditional probabilities.

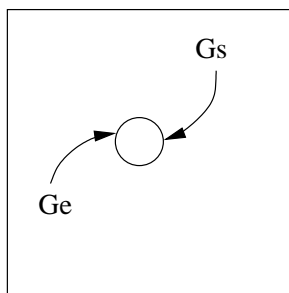
Recall that the conditional probability of  $A$  given  $E$  is:

$$P(A|E) = \frac{P(A \cap E)}{P(E)}$$

So, the likelihood of  $H_1$  is:

$$\begin{aligned} \mathcal{L}(H_1) &= P(G_s = G | G_e = G \text{ and } \mathcal{S} = \mathcal{E}) \\ &= \frac{P(G_s \cap G_e)}{P(G_e)} = \frac{P(G_s)}{P(G_e)} = \frac{P(G)}{P(G)} = 1.0 \end{aligned}$$

This is depicted below where the box represents all possible samples of DNA material and the circle corresponds to the DNA of the suspect and of the evidence:



- E. The likelihood of  $H_0$

1.  $\mathcal{L}(H_0) =$  probability that the suspect will have DNA =  $G$  **given** that the evidence DNA =  $G$  and (hypothesis)  $\mathcal{S} \neq \mathcal{E}$  (suspect did not leave the evidence).
2. This probability is not so obvious; so, we will use the definition of conditional probability to calculate:

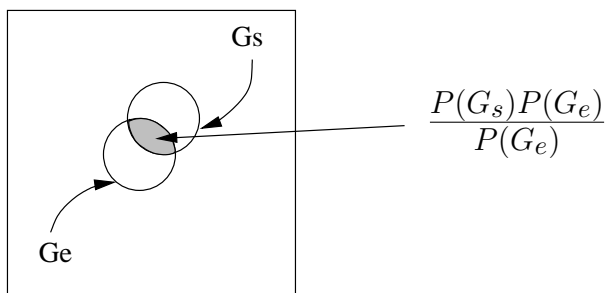
$$\begin{aligned} \mathcal{L}(H_0) &= P(G_e = G | G_e = G \text{ and } \mathcal{S} \neq \mathcal{E}) \\ &= \frac{P(G_e \cap G_e)}{P(G_e)} \\ &= \frac{P(G_s)P(G_e)}{P(G_e)} = P(G) \end{aligned}$$

We multiply the probabilities in the last step because we assume that the events  $G_s$  and  $G_e$  are **independent**. In other words,  $\mathcal{S} \neq \mathcal{E}$ . (QUIZ: What about twins?)

## DNA Fingerprinting Probabilities

---

3. Here is the picture of this situation.



Now we have two hypotheses and two likelihoods. Which is the most likely?

F.  $\mathcal{LR}$  for the fingerprint

1. The BOTTOM LINE:

$$\mathcal{LR} = \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \frac{1}{P(G)}$$

2. By proper choice of the loci used to observe the VNTR (Variable Number of Tandem Repeats), we can make  $P(G)$  very, very small. This will make  $\mathcal{LR}$  very, very large (if  $\mathcal{S} = \mathcal{E}$ ).

E.g., if  $P(G) = 10^{-12}$ , then  $\mathcal{LR} = 10^{12}$ .

3. How does one explain this to the jury?

**Correct:** The evidence (i.e.,  $G_e = G_s$ , or the suspect's DNA is  $G$ ) is  $10^{12}$  times more likely to have been observed if the crime scene material was left by the suspect ( $\mathcal{S}$ ) than if it had been left by another person.

**Incorrect:** If  $\mathcal{LR}$  is large, the odds are overwhelming that one person (the suspect) was the source of both samples ( $G_s$  and  $G_e$ ).

**Incorrect:** The probability of observing  $G_e$  in a random draw from the population is  $1/\mathcal{LR} = 10^{-12}$ .

Or: Only 1 person in  $10^{12}$  will have  $G_e$ .

4. Testing the significance of  $\mathcal{LR}$

i. Definition: *the ratio of the likelihoods of two competing hypotheses*

ii. The die example:

$$\mathcal{LR} = \frac{\mathcal{L}(\theta = 0.4)}{\mathcal{L}(\theta = 0.1667)} = \frac{0.346}{0.1608} = 2.15$$

iii. In words: "It is 2.15 times more likely that the observed sample was produced by a die with  $\theta = 0.4$  than by a die with  $\theta = 0.1667$ ."

iv. But, is 2.15 a large number?

Can not tell by looking, but probably NOT

Rule of Thumb:  $\mathcal{LR} > 10 \Rightarrow$  significant difference

Can do better with statistical tests, e.g.,  $\chi^2$

G. Calculating  $P(G)$

1. The success of the method depends on good estimates of  $P(G)$ .

2. Assumptions for VNTR analysis

i. Alleles are independent

ii. Loci are independent

iii. Frequencies of alleles and loci are homogeneous in the population

## DNA Fingerprinting Probabilities

---

- iv. Two parents
- 3. Calculations
  - i. Multiply allele frequencies within a locus times 2 (for both parents) to get the probability of the locus genotype
  - ii. Multiply locus probabilities across loci
  - iii. Use the allele frequencies for the population at large (not ethnic groups)
- 4. Example: 2 subpopulations of A and B.

loci	Sub-Population							
	A				B			
	1		2		1		2	
alleles	a	b	a	b	a	b	a	b
frequency	0.01	0.05	0.02	0.03	0.005	0.007	0.02	0.005

For each subpopulation the values are:

$$\begin{aligned}
 P(G_A) &= 2(0.01)(0.05) \times 2(0.02)(0.03) \\
 &= 1.2 \times 10^{-6}
 \end{aligned}$$

$$\begin{aligned}
 P(G_B) &= 2(0.005)(0.007) \times 2(0.02)(0.005) \\
 &= 1.4 \times 10^{-8}
 \end{aligned}$$

Assume total population is just the average of A and B for allele frequencies

TOTAL			
1		2	
a	b	a	b
0.0075	0.0285	0.02	0.0175

Compare:

$$\begin{aligned}
 \mathcal{LR}(A) &= 8.33 \times 10^5 \\
 \mathcal{LR}(B) &= 7.14 \times 10^7 \\
 \mathcal{LR}(Total) &= 3.34 \times 10^6
 \end{aligned}$$

- i. Does the difference among the above calculations really matter for ascribing guilt or innocence?
- ii. What would happen to the difference if Subpopulation B was only 1/5 as numerically abundant as Subpopulation A?
- iii. If we have more loci (e.g, 8, which is more realistic), the importance of the differences are reduced further. Assume the frequencies in 6 more loci are the same as the first 2 for A, B, and Total; then:

$$\begin{aligned}
 \mathcal{LR}(A) &= 4.8 \times 10^{23} \\
 \mathcal{LR}(B) &= 2.6 \times 10^{31} \\
 \mathcal{LR}(Total) &= 1.2 \times 10^{26}
 \end{aligned}$$

These are so large that it doesn't matter which population we use.

- iv. Even if we wanted to separate the sub-populations, how should we determine the sub-groups? By race?, national origin?, physical characteristics?, economic status?, dress code? The question raises important social issues. Moreover, a computer study in which individuals were **randomly** assigned to arbitrary groups produced as much heterogeneity in VNTR patterns (differences among groups) as obtained using the race of the individuals.

## DNA Fingerprinting Probabilities

---

### VI. References

Roeder, K. 1994 DNA fingerprinting: a review of the controversy. *Statistical Science*. 9(2):222-247.

Roeder, K. 1994 Rejoinders to: DNA fingerprinting: a review of the controversy. *Statistical Science*. 9(2):262-278.

Weir, B.S. 1992 Population genetics in the forensic DNA debate. *Proceedings National Academy of Science, USA*. 89:11654-11659.

Weir, B. 1995 DNA statistics in the Simpson matter. *Nature Genetics*. 11:365-368.