ALLELIC INDEPENDENCE

Probability

Three possible definitions for probability:

- Equiprobable outcomes definition.
- Long-run frequency definition.
- Subjective probability.

Axioms of Probability

1. $0 \leq \Pr(A) \leq 1, \Pr(A|A) = 1.$

2. $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$ if A, B mutually exclusive.

3. $\Pr(A \text{ and } B) = \Pr(A) \Pr(B|A).$

Law of Total Probability

For any event E and any set of mutually exclusive and exhaustive events $\{S_i\}$:

$$\Pr(E) = \sum_i \Pr(E|S_i) \Pr(S_i)$$

Bayes' Theorem

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

Working Group Bayesian Exercise

A rapid test for covid-19 is set up outside a supermarket and is available to anyone who wishes. The test has a false-positive rate of 5% and a false-negative rate of 30%.

If the disease has a prevalence in that population of 20%. What is the probability a person who tests positive does actually have the disease? i.e. calculate $\Pr(B|A)$ if A is the event that a test is positive, and B is the event that a person has the disease. Use Bayes' Theorem.

Single Crime Scene Stain

Suppose a blood stain is found at a crime scene, and it must have come from the offender. A suspect is identified and provides a blood sample. The crime scene sample and the suspect have the same (DNA) "type."

The prosecution subsequently puts to the court the proposition (or hypothesis or explanation):

H_p: The suspect left the crime stain.

The symbol H_p is just to assist in the formal analysis. It need not be given in court.

Transfer Evidence Notation

G_S, G_C are the DNA types for suspect and crime sample.
$G_S = G_C$.
I is non-DNA evidence.

Before the DNA typing, probability of H_p is conditioned on I.

After the typing, probability of H_p is conditioned on G_S, G_C, I.

Updating Uncertainty

Method of updating uncertainty, or changing $\Pr(\text{Hypothesis}_p)$ to $\Pr(\text{Hypothesis}_p|\text{Evidence})$ uses Bayes' theorem:

$$\Pr(\text{Hypothesis}_p|\text{Evidence}) = \frac{\Pr(\text{Evidence}|\text{Hypothesis}_p)\Pr(\text{Hypothesis}_p)}{\Pr(\text{Evidence})}$$

We can't evaluate $\Pr(\text{Evidence})$ without additional information, and we don't know $\Pr(\text{Hypothesis}_p)$.

Can proceed by introducing alternative to Hypothesis_p.

First Principle of Evidence Interpretation

To evaluate the uncertainty of a proposition, it is necessary to consider at least one alternative proposition.

The simplest alternative explanation for a single stain is:

H_d: Some other person left the crime stain.

Evett IW, Weir BS. 1998. "Interpreting DNA Evidence."

Can be downloaded from:

`www.biostat.washington.edu/~bsweir/InterpretingDNAEvidence`

Updating Odds

From the odds form of Bayes' theorem:

$$\frac{\text{Pr}(\text{Hypothesis}_p|\text{Evidence})}{\text{Pr}(\text{Hypothesis}_d|\text{Evidence})} = \frac{\text{Pr}(\text{Evidence}|\text{Hypothesis}_p)}{\text{Pr}(\text{Evidence}|\text{Hypothesis}_d)} \times \frac{\text{Pr}(\text{Hypothesis}_p)}{\text{Pr}(\text{Hypothesis}_d)}$$

i.e. Posterior odds = LR $\times$ Prior odds

where

$$\text{LR} = \frac{\text{Pr}(\text{Evidence}|\text{Hypothesis}_p)}{\text{Pr}(\text{Evidence}|\text{Hypothesis}_d)}$$

Questions for a Court to Consider

The trier of fact needs to address questions of the kind

- What is the probability that the prosecution proposition is true given the evidence,
  $\Pr(H_p|G_C, G_S, I)$?

- What is the probability that the defense proposition is true given the evidence,
  $\Pr(H_d|G_C, G_S, I)$?

Questions for Forensic Scientist to Consider

The forensic scientist must address different questions:

- What is the probability of the DNA evidence if the prosecution proposition is true,
  $\Pr(G_C, G_S | H_p, I)$?

- What is the probability of the DNA evidence if the defense proposition is true,
  $\Pr(G_C, G_S | H_d, I)$?

Important to articulate H_p, H_d. Also important not to confuse the difference between these two sets of questions.

Second Principle of Evidence Interpretation

Evidence interpretation is based on questions of the kind 'What is the probability of the evidence given the proposition.'

This question is answered for alternative explanations, and the ratio of the probabilities presented. It is not necessary to use the words "likelihood ratio". Use phrases such as:
'The probability that the crime scene DNA type is the same as the suspect's DNA type is one million times higher if the suspect left the crime sample than if someone else left the sample.'

Third Principle of Evidence Interpretation

Evidence interpretation is conditioned not only on the alternative propositions, but also on the framework of circumstances within which they are to be evaluated.

The circumstances may simply be the population to which the offender belongs so that probabilities can be calculated. Forensic scientists must be clear in court about the nature of the non-DNA evidence I, as it appeared to them when they made their assessment. If the court has a different view then the scientist must review the interpretation of the evidence.

Example

"In the analysis of the results I carried out I considered two alternatives: either that the blood samples originated from Pengelly or that the ... blood was from another individual. I find that the results I obtained were at least 12,450 times more likely to have occurred if the blood had originated from Pengelly than if it had originated from someone else."

Example

Question: "Can you express that in another way?"

Answer: "It could also be said that 1 in 12,450 people would have the same profile ... and that Pengelly was included in that number ... very strongly suggests the premise that the two blood stains examined came from Pengelly."

[Testimony of M. Lawton in R. v Pengelly 1 NZLR 545 (CA), quoted by
Robertson B, Vignaux GA, Berger CEH. 2016.*Interpreting Evidence (Second Edition)*. Wiley.

Likelihood Ratio

$$\mathsf{LR} \;=\; \frac{\mathsf{Pr}(G_C, G_S | H_p, I)}{\mathsf{Pr}(G_C, G_S | H_d, I)}$$

Apply laws of probability to change this into

$$\mathsf{LR} \;=\; \frac{\mathsf{Pr}(G_C | G_S, H_p, I)\,\mathsf{Pr}(G_S | H_p, I)}{\mathsf{Pr}(G_C | G_S, H_d, I)\,\mathsf{Pr}(G_S | H_d, I)}$$

Likelihood Ratio

Whether or not the suspect left the crime sample (i.e. whether or not H_p or H_d is true) provides no information about his genotype:

$$\Pr(G_S|H_p, I) = \Pr(G_S|H_d, I) \;=\; \Pr(G_S|I)$$

so that

$$\mathsf{LR} \;=\; \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)}$$

This is the form that allows the consideration of relatives and/or population structure, as well as drop-out and drop-in.

Likelihood Ratio

$$\text{LR} = \frac{\Pr(G_C|G_S, H_p, I)}{\Pr(G_C|G_S, H_d, I)}$$

When $G_C = G_S$, and when they are for the same person (H_p is true):

$$\Pr(G_C|G_S, H_p, I) = 1$$

so the likelihood ratio becomes

$$\text{LR} = \frac{1}{\Pr(G_C|G_S, H_d, I)}$$

This is the reciprocal of the probability of the *match probability*, the probability of profile G_C, conditioned on having seen profile G_S in a different person (i.e. H_d) and on I.

Likelihood Ratio

$$\mathsf{LR} \;=\; \frac{1}{\Pr(G_C|G_S, H_d, I)}$$

The next step depends on the circumstances I. If these say that knowledge of the suspect's type does not affect our uncertainty about the offender's type when they are different people (i.e. when H_d is true):

$$\Pr(G_C|G_S, H_d, I) \;=\; \Pr(G_C|H_d, I)$$

and then likelihood ratio becomes

$$\mathsf{LR} \;=\; \frac{1}{\Pr(G_C|H_d, I)}$$

The LR is now the reciprocal of the *profile probability* of profile G_C.

Profile and Match Probabilities

Dropping mention of the other information I, the quantity $\Pr(G_C)$ is the probability that a person randomly chosen from a population will have profile type G_C. This profile probability usually very small and, although it is interesting, it is not the most relevant quantity.

Of relevance is the match probability, the probability of seeing the profile in a randomly chosen person after we have already seen that profile in a typed person (the suspect). The match probability is bigger than the profile probability. Having seen a profile once there is an increased chance we will see it again. This is the genetic essence of DNA evidence.

Likelihood Ratio

The estimated probability in the denominator of LR is determined on the basis of judgment, informed by I. Therefore the nature of I (as it appeared to the forensic scientist at the time of analysis) must be explained in court along with the value of LR. If the court has a different view of I, then the scientist will need to review the interpretation of the DNA evidence.

Meaning of Likelihood Ratios

There is a personal element to interpreting DNA evidence, and there is no "right" value for the LR. (There is a right answer to the question of whether the suspect left the crime stain, but that is not for the forensic scientist to decide.)

The denominator for LR is conditioned on the stain coming from an unknown person, and "unknown" may be hard to define. A relative? Someone in that town? Someone in the same ethnic group? (What is an ethnic group?)

Interpretation of LR

"The likelihood ratio forms the basis for optimal decision rules regarding the competing hypotheses. A likelihood ratio greater than 1 supports H_p, while a likelihood ratio less than 1 supports H_d. A large likelihood ratio does however not necessarily imply that the posterior odds are high, because the prior odds also play a role. Moreover, high posterior odds do not necessarily imply that H_p is likely, since the odds only compare just two hypotheses, while there may exist another hypothesis that is in fact more likely."

Kruijver M, et al. 2015. Forensic Science International: Genetics 16:226-231.

When is an LR large?

For a single-contributor evidence profile, completely matched by the suspect's profile, the LR is LR $= 1/P_E$ where P_E is the probability a person drawn randomly from the population has the same profile as the evidence.

For all possible profiles at this set of loci, whose probabilities $(1/x)$ sum to 1,

$$P_{\mathsf{Min}} \leq \cdots \leq P_E \cdots \leq (1/x) \leq \cdots \leq P_{\mathsf{Max}}$$

$$1/P_{\mathsf{Min}} \geq \cdots \mathsf{LR} \cdots \geq x \geq \cdots \geq 1/P_{\mathsf{Max}}$$

So the probability that LR is at least equal to x is no more than $1/x$. The size therefore depends on the profile and the population (represented by the database). This extends to mixtures, relatives etc.

Testing for Allelic Independence

What is the probability a person has a particular DNA profile? What is the probability a person has a particular profile if it has already been seen once?

The first question is a little easier to think about, but difficult to answer in practice: it is very unlikely that a profile will be seen in any sample of profiles. Even for one STR locus with 10 alleles, there are 55 different genotypes and most of those will not occur in a sample of a few hundred profiles.

For locus D3S1358 in the African American population, the FBI frequency database shows that 31 of the 55 genotype counts are zero. Estimating the population frequencies for these 31 types as zero doesn't seem sensible.

D3S1358 Genotype Counts

Observed	<12	12	13	14	15	16	17	18	19	>19
<12	0									
12	0	0								
13	0	0	0							
14	0	0	0	2						
15	0	0	1	19	15					
16	1	1	1	15	39	19				
17	0	0	2	10	26	24	9			
18	1	0	1	2	6	10	3	0		
19	0	0	0	1	0	0	1	0	0	
>19	0	0	0	0	1	0	0	0	0	0

The number in row i and column j is the observed count of indivuals with alleles i and j.

Hardy-Weinberg Law

A solution to the problem is to assume that the Hardy-Weinberg Law holds. For a random mating population, expect that genotype frequencies are products of allele frequencies.

For a locus with two alleles, A, a:

$$\begin{aligned} P_{AA} &= (p_A)^2 \\ P_{Aa} &= 2p_A p_a \\ P_{aa} &= (p_a)^2 \end{aligned}$$

For a locus with several alleles A_i:

$$\begin{aligned} P_{A_iA_i} &= (p_{A_i})^2 \\ P_{A_iA_j} &= 2p_{A_i}p_{A_j} \end{aligned}$$

D3S1358 Hardy-Weinberg Calculations

The allele counts for D3S1358 in the African-American sample are:

										Total
Allele	<12	12	13	14	15	16	17	18	19	>19	
Count	2	1	5	51	122	129	84	23	2	1	420

If the Hardy-Weinberg Law holds, then we would expect to see $n\tilde{p}_{13}^2 = 210 \times (5/420)^2 = 0.03$ individuals of type 13,13 in a sample of 210 individuals.

Also, we would expect to see $2n\tilde{p}_{13}\tilde{p}_{14} = 420 \times (5/420) \times (51/420) = 0.61$ individuals of type 13,14 in a sample of 210 individuals.

Other values are shown on the next slide.

D3S1358 Observed and Expected Counts

		<12	12	13	14	15	16	17	18	19	>19
<12	Obs.	0									
	Exp.	0.0									
12	Obs.	0	0								
	Exp.	0.0	0.0								
13	Obs.	0	0	0							
	Exp.	0.0	0.0	0.0							
14	Obs.	0	0	0	2						
	Exp.	0.2	0.1	0.6	3.1						
15	Obs.	0	0	1	19	15					
	Exp.	0.6	0.3	1.5	14.8	17.7					
16	Obs.	1	1	1	15	39	19				
	Exp.	0.6	0.3	1.5	15.7	37.5	19.8				
17	Obs.	0	0	2	10	26	24	9			
	Exp.	0.4	0.2	1.0	10.2	24.4	25.8	8.4			
18	Obs.	1	0	1	2	6	10	3	0		
	Exp.	0.1	0.1	0.3	2.8	6.7	7.1	4.6	0.6		
19	Obs.	0	0	0	1	0	0	1	0	0	
	Exp.	0.0	0.0	0.0	0.2	0.6	0.6	0.4	0.1	0.0	
>19	Obs.	0	0	0	0	1	0	0	0	0	0
	Exp.	0.0	0.0	0.0	0.1	0.3	0.3	0.2	0.1	0.0	0.0

Testing for Hardy-Weinberg Equilibrium

A test of the Hardy-Weinberg Law will somehow decide if the observed and expected numbers are sufficiently similar that we can proceed as though the law can be used.

In one of the first applications of Hardy-Weinberg testing in a US forensic setting:

> "To justify applying the classical formulas of population genetics in the Castro case the Hispanic population must be in Hardy-Weinberg equilibrium. Applying this test to the Hispanic sample, one finds spectacular deviations from Hardy-Weinberg equilibrium."

Lander ES. 1989. DNA fingerprinting on trial. Nature 339: 501-505.

NIST Database

The NIST database shows p-values for Hardy-Weinberg tests.