PROBABILITY THEORY

Probability Theory

We wish to attach probabilities to different kinds of events (or hypotheses or propositions):

- Event A: the next card is an Ace.
- Event R: it will rain tomorrow.
- Event C: the suspect left the crime stain.

Probabilities

Assign probabilities to events: Pr(A) or p_A or even p means "the probability that event A is true." All probabilities are conditional on some information I, so should write Pr($A|I$) for "the probability that A is true given that I is known."

No matter how probabilities are defined, they need to follow some mathematical laws in order to lead to consistent theories.

First Law of Probability

$$0 \leq \Pr(A|I) \leq 1$$

$$\Pr(A|A, I) = 1$$

If A is the event that a die shows an even face (2, 4, or 6), what is I? What is $\Pr(A|I)$?

Second Law of Probability

If A, B are mutually exclusive given I

$$\Pr(A \text{ or } B|I) \;=\; \Pr(A|I) + \Pr(B|I)$$

$$\text{so } \Pr(\bar{A}|I) \;=\; 1 - \Pr(A|I)$$

($\bar{A}$ means not-A).

If A is the event that a die shows an even face, and B is the event that the die shows a 1, verify the Second Law.

Third Law of Probability

$$\Pr(A \text{ and } B|I) \;=\; \Pr(A|B,I) \times \Pr(B|I)$$

If A is event that die shows an even face, and B is the event that the die shows a 1, verify the Third Law.

Will generally omit the I from now on.

Independent Events

Events A and B are independent if knowledge of one does not affect probability of the other:

$$\begin{aligned} \Pr(A|B) &= \Pr(A) \\ \Pr(B|A) &= \Pr(B) \end{aligned}$$

Therefore, for independent events

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

This may be written as

$$\Pr(AB) = \Pr(A)\Pr(B)$$

Law of Total Probability

Because B and $\bar{B}$ are mutually exclusive and exhaustive:

$$\Pr(A) \;=\; \Pr(A|B)\Pr(B) + \Pr(A|\bar{B})\Pr(\bar{B})$$

If A is the event that die shows a 3, B is the event that the die shows an even face, and $\bar{B}$ the event that the die shows an odd face, verify the Law of Total Probability.

Odds

The odds $O(A)$ of an event A are the probability of the event being true divided by the probability of the event not being true:

$$O(A) = \frac{\Pr(A)}{\Pr(\bar{A})}$$

This can be rearranged to give

$$\Pr(A) = \frac{O(A)}{1 + O(A)}$$

Odds of 10 to 1 are equivalent to a probability of 10/11.

Bayes' Theorem

The third law of probability can be used twice to reverse the order of conditioning:

$$\begin{aligned} \Pr(B|A) &= \frac{\Pr(B \text{ and } A)}{\Pr(A)} \\ &= \frac{\Pr(A|B)\Pr(B)}{\Pr(A)} \end{aligned}$$

Odds Form of Bayes' Theorem

From the third law of probability

$$\begin{aligned} \Pr(B|A) &= \Pr(A|B)\Pr(B)/\Pr(A) \\ \Pr(\bar{B}|A) &= \Pr(A|\bar{B})\Pr(\bar{B})/\Pr(A) \end{aligned}$$

Taking the ratio of these two equations:

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

Posterior odds = likelihood ratio $\times$ prior odds.

Birthday Problem

Forensic scientists in Arizona looked at the 65,493 profiles in the Arizona database and reported that two profiles matched at 9 loci out of 13. They reported a "match probability" for those 9 loci of 1 in 754 million. Are the numbers 65,493 and 754 million inconsistent?

Troyer et al., 2001. Proc Promega 12th Int Symp Human Identification.

To begin to answer this question suppose that every possible profile has the same profile probability P and that there are N profiles in a database (or in a population). The probability of at least one pair of matching profiles in the database is one minus the probability of no matches.

Birthday Problem

Choose profile 1. The probability that profile 2 does not match profile 1 is $(1-P)$. The probability that profile 3 does not match profiles 1 or 2 is $(1-2P)$, etc. So, the probability P_M of at least one matching pair is

$$P_M = 1-\{1(1-P)(1-2P)\cdots[1-(N-1)P]\}$$

$$\approx 1-\prod_{i=0}^{N-1} e^{-iP} \approx 1-e^{-N^2P/2}$$

If $P=1/365$ and $N=23$, then $P_M=0.51$. So, approximately, in a room of 23 people there is greater than a 50% probability that two people have the same birthday.

Birthday Problem

If $P = 1/(754 \text{ million})$ and $N = 65,493$, then $P_M = 0.98$ so it is highly probable there would be a match. There are other issues, having to do with the four non-matching loci, and the possible presence of relatives in the database.

If $P = 10^{-16}$ and $N = 300$ million, then $P_M =$ is essentially 1. It is almost certain that two people in the US have the same rare DNA profile.

Statistics

- Probability: For a given model, what do we expect to see?

- Statistics: For some given data, what can we say about the model?

- Example: A marker has an allele A with frequency p_A.

  - Probability question: If $p_A = 0.5$, and if alleles are independent, what is the probability of AA?

  - Statistics question: If a sample of 100 individuals has 23 AA's, 48 Aa's and 29 aa's, what is an estimate of p_A?

LIKELIHOOD RATIOS

Transfer Evidence

Relevant Evidence

Rule 401 of the US Federal Rules of Evidence:

"Relevant evidence" means evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.

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

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

Apply laws of probability to change this into

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

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

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

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

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

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

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

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.

Random Samples

The circumstances I may define a population or racial group. The probability is estimated on the basis of a sample from that population.

When we talk about DNA types, by "selecting a person at random" we mean choosing a person in such a way as to be as uncertain as possible about their DNA type.

Convenience Samples

The problem with a formal approach is that of defining the population: if we mean the population of a town, do we mean *every* person in the town at the time the crime was committed? Do we mean some particular area of the town? One sex? Some age range?

It seems satisfactory instead to use a convenience sample, i.e. a set of people from whom it is easy to collect biological material in order to determine their DNA profiles. These people are not a random sample of people, but they have not been selected on the basis of their DNA profiles.

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?)

Meaning of Frequencies

What is meant by "the frequency of the matching profile is 1 in 57 billion"?

It is an estimated probability, obtained by multiplying together the allele frequencies, and refers to an infinite random mating population. It has nothing to do with the size of the world's population.

The question is really whether we would see the profile in two people, given that we have already seen it in one person. This conditional probability may be very low, but has nothing to do with the size of the population.

Explaining Likelihood Ratios

"There is a broad scientific consensus that likelihoods are the primary tool for DNA evidence evaluation and that forensic experts should present the strength of DNA evidence using the likelihood ratio. There are good reasons to do so: the likelihood ratio measures the strength of the evidence and the likelihood ratio approach allows a finder of fact to combine the evidence with background information or other evidence in a coherent way. Moreover, a desirable feature of the likelihood ratio approach is that the inferential process is out of the hands of the forensic expert, but rests with the court instead."

Kruijver M, Meester R, Slooten R. FSI:Genetics 16:221-231 (2015).

Explaining Likelihood Ratios

"One of the main arguments against LR is that it may be difficult to explain the meaning of a large likelihood ratio in court. The RMNE approach has a more intuitive understanding in the sense that it presents the probability of not excluding a random man as a contributor to the evidence, but the main criticism against its use is that it wastes information."

Dorum G, et al. FSI:Genetics 9:93-1-1 (2014).

"Despite the broad scientific support for the likelihood ratio approach, there has been resistance against its use in the courtroom. The main critique on the approach is that it would be difficult to explain the meaning of a likelihood ratio. In response to this critique, several authors have recently suggested to interpret the likelihood ratio using an associated p-value."

Kruijver M, Meester R, Slooten R. FSI:Genetics 16:221-231 (2015).

Distribution of LRs

The idea behind LR p-values was to convey some sense of the size of the calculated LR. Is one million big? What about one billion? What about 100?

For evidence E and hypotheses H_p, H_d, the LR is

$$\text{LR} = \frac{\Pr(E|H_p)}{\Pr(E|H_d)}$$

The simplest case is for H_p says Suspect S is a contributor to E, and H_d says S is not a contributor. For a single-contributor stain, and no drop-out of drop-in, only people with the same genotype as E would give a non-zero LR and all those people would have the same LR.

In more complicated cases, a question is whether some person other than S would have a higher LR.

Distribution of LRs

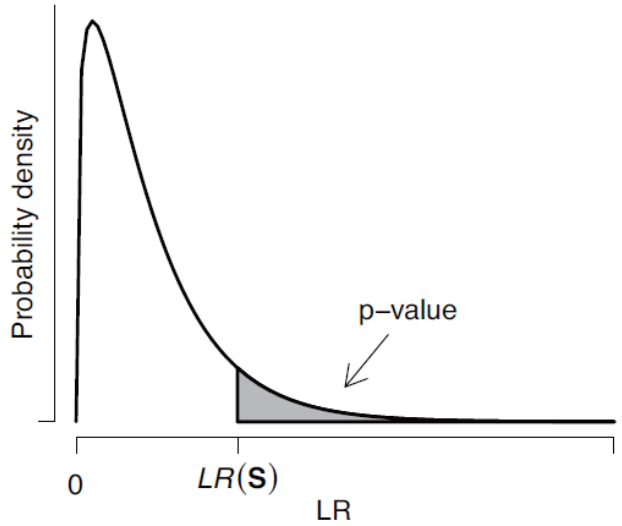
Now write

$$\text{LR}(S) = \frac{\Pr(E|H_p : S \text{ is a contributor to } E)}{\Pr(E|H_d : S \text{ is not a contributor to } E))}$$

For some person R other than S:

$$\text{LR}(R) = \frac{\Pr(E|H_p : R \text{ is a contributor to } E)}{\Pr(E|H_d : S \text{ is not a contributor to } E))}$$

The denominator has not changed. This has been called the "H_d-true LR." If R is a random person other than S, then the average of all values of LR(R) equals one. If all possible values of LR(R) are plotted:



LR p values

The p-value for LR(S) is

$$p \;=\; \Pr[\mathrm{LR}(R) > \mathrm{LR}(S)]$$

and it can be shown that this p-value is less than $1/\mathrm{LR}(S)$. Large LR values necessarily have small p-values, so the p-value is not helpful.

"How (un)likely it is to obtain a piece of evidence is, however, not relevant for drawing inferences about the two competing hypotheses in the likelihood ratio approach. ... it will always be the case that the p-value numerically suggests stronger evidence than the likelihood ratio, and this is an indication already that p-values can be misleading."

Kruijver M, Meester R, Slooten R. FSI:Genetics 16:221-231 (2015).