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_1: The suspect left the crime stain.

The symbol H_1 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_1 is conditioned on I.

After the typing, probability of H_1 is conditioned on G_S, G_C, I.

Updating Uncertainty

Method of updating uncertainty, or changing $\Pr(\text{Hypothesis}_1)$ to $\Pr(\text{Hypothesis}_1|\text{Evidence})$ uses Bayes' theorem:

$$\Pr(\text{Hypothesis}_1|\text{Evidence}) = \frac{\Pr(\text{Evidence}|\text{Hypothesis}_1)\Pr(\text{Hypothesis}_1)}{\Pr(\text{Evidence})}$$

We can't evaluate $\Pr(\text{Evidence})$ without additional information, and we don't know $\Pr(\text{Hypothesis}_1)$.

Can proceed by introducing alternative to Hypothesis_1.

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_2: 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:

$$\frac{\Pr(\text{Hypothesis}_1|\text{Evidence})}{\Pr(\text{Hypothesis}_2|\text{Evidence})} = \frac{\Pr(\text{Evidence}|\text{Hypothesis}_1)}{\Pr(\text{Evidence}|\text{Hypothesis}_2)} \times \frac{\Pr(\text{Hypothesis}_1)}{\Pr(\text{Hypothesis}_2)}$$

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

where

$$\text{LR} = \frac{\Pr(\text{Evidence}|\text{Hypothesis}_1)}{\Pr(\text{Evidence}|\text{Hypothesis}_2)}$$

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,
  $\Pr(H_1|G_C, G_S, I)$?

- What is the probability that the defense proposition is true given the evidence,
  $\Pr(H_2|G_C, G_S, I)$?

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,
  $\Pr(G_C, G_S | H_1, I)$?

- What is the probability of the DNA evidence if the defense proposition is true,
  $\Pr(G_C, G_S | H_2, I)$?

Important to articulate H_1, H_2. 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

$$\text{LR} = \frac{\Pr(G_C, G_S | H_1, I)}{\Pr(G_C, G_S | H_2, I)}$$

Apply laws of probability to change this into

$$\text{LR} = \frac{\Pr(G_C | G_S, H_1, I) \Pr(G_S | H_1, I)}{\Pr(G_C | G_S, H_2, I) \Pr(G_S | H_2, I)}$$

Likelihood Ratio

Whether or not the suspect left the crime sample (i.e. whether or not H_1 or H_2 is true) provides no information about his genotype:

$$\Pr(G_S|H_1, I) = \Pr(G_S|H_2, I) \;=\; \Pr(G_S|I)$$

so that

$$\mathrm{LR} \;=\; \frac{\Pr(G_C|G_S, H_1, I)}{\Pr(G_C|G_S, H_2, I)}$$

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

Likelihood Ratio

$$\text{LR} = \frac{\Pr(G_C|G_S, H_1, I)}{\Pr(G_C|G_S, H_2, I)}$$

When $G_C = G_S$, and when they are for the same person (H_1 is true):

$$\Pr(G_C|G_S, H_1, I) = 1$$

so the likelihood ratio becomes

$$\text{LR} = \frac{1}{\Pr(G_C|G_S, H_2, I)}$$

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_2) and on I.

Likelihood Ratio

$$\text{LR} \;=\; \frac{1}{\Pr(G_C|G_S, H_2, I)}$$

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_2 is true):

$$\Pr(G_C|G_S, H_2, I) \;=\; \Pr(G_C|H_2, I)$$

and then likelihood ratio becomes

$$\text{LR} \;=\; \frac{1}{\Pr(G_C|H_2, I)}$$

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.

ALLELIC INDEPENDENCE

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.

VNTR "Coalescence"

Forensic DNA profiling initially used minisatellites, or VNTR loci, with large numbers of alleles. Heterozygotes would be scored as homozygotes if the two alleles were so similar in length that they coalesced into one band on an autoradiogram. Small alleles often not detected at all, and this is a likely cause of Lander's finding (Devlin et al, Science 249:1416-1420.) .

Considerable debate in early 1990s on alternative "binning" strategies for reducing the number of alleles (Science 253:1037-1041, 1991).

Typing has moved to microsatellites with fewer and more easily distinguished alleles, but testing for Hardy-Weinberg equilibrium continues. There are still reasons why the law may not hold.

Population Structure can Cause Departure from HWE

If a population consists of a number of subpopulations, each in HWE but with different allele frequencies, there will be a departure from HWE at the population level. This is the Wahlund effect.

Suppose there are two equal-sized subpopulations, each in HWE but with different allele frequencies, then

	Subpopn 1	Subpopn 2	Total Popn
p_A	0.6	0.4	0.5
p_a	0.4	0.6	0.5
P_{AA}	0.36	0.16	$0.26 > (0.5)^2$
P_{Aa}	0.48	0.48	$0.48 < 2(0.5)(0.5)$
P_{aa}	0.16	0.36	$0.26 > (0.5)^2$

Population Structure

Effect of population structure taken into account with the "theta-correction." Matching probabilities allow for a variance in allele frequencies among subpopulations.

$$\Pr(AA|AA) = \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}$$

where p_A is the average allele frequency over all subpopulations. We will come back to this expression.

Population Admixture

A population might represent the recent admixture of two parental populations. With the same two populations as before but now with 1/4 of marriages within population 1, 1/2 of marriages between populations 1 and 2, and 1/4 of marriages within population 2. If children with one or two parents in population 1 are considered as belonging to population 1, there is an excess of heterozygosity in the offspring population.

If the proportions of marriages within populations 1 and 2 are both 25% and the proportion between populations 1 and 2 is 50%, the next generation has

	Population 1	Population 2
P_{AA}	$0.09 + 0.12 = 0.21$	0.04
P_{Aa}	$0.12 + 0.26 = 0.38$	0.12
P_{aa}	$0.04 + 0.12 = 0.16$	0.09
	0.75	0.25

Exact HWE Test

The preferred test for HWE is an "exact" one. The test uses the conditional probability of the genotypic counts (n_{AA}, n_{Aa}, n_{aa}) given the allelic counts (n_A, n_a) and given HWE:

$$\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a, \text{HWE}) = \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!} \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

Reject the Hardy-Weinberg hypothesis if this probability is unusually small.

Exact HWE Test Example

Reject the HWE hypothesis if the probability of the genotypic array, conditional on the allelic array, is among the smallest probabilities for all the possible sets of genotypic counts for those allele counts.

As an example, consider $(n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49)$. The allele counts are $(n_A = 2, n_a = 98)$ and there are only two possible genotype arrays:

AA	Aa	aa	$\Pr(n_{AA}, n_{Aa}, n_{aa} \mid n_A, n_a, \text{HWE})$
1	0	49	$\frac{50!}{1!0!49!}\frac{2^0 2!98!}{100!} = \frac{1}{99}$
0	2	48	$\frac{50!}{0!2!48!}\frac{2^2 2!98!}{100!} = \frac{98}{99}$

Exact HWE Test

The probability of the data on the previous slide, conditional on the allele frequencies and on HWE, is $1/99 = 0.01$. This is less than the conventional 5% significance level.

In general, the p-value is the (conditional) probability of the data plus the probabilities of all the less-probable datasets. The probabilities are all calculated assuming HWE is true.

Permutation Test

For large sample sizes and many alleles per locus, there are too many genotypic arrays for a complete enumeration and a determination of which are the least probable 5% arrays.

A large number of the possible arrays is generated by permuting the alleles among genotypes, and calculating the proportion of these permuted genotypic arrays that have a smaller conditional probability than the original data. If this proportion is small, the Hardy-Weinberg hypothesis is rejected.

Permutation Test

Fold a piece of paper into five horizonal strips. Mark each strip with two alleles to represent five genotypes:

Card 1:	A	A
Card 2:	A	A
Card 3:	A	A
Card 4:	a	a
Card 5:	a	a

Tear the cards in half to give a deck of 10 cards, each with one allele. Shuffle the deck and deal into 5 pairs, to give five genotypes.

Permutation Test

The permuted set of genotypes fall into one of four types:

AA	Aa	aa	Number of times
3	0	2	
2	2	1	
1	4	0	

Permutation Test

Check the following theoretical values for the proportions of each of the three types, from the expression:

$$\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} \times \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

AA	Aa	aa	Conditional Probability
3	0	2	$\frac{1}{21} = 0.048$
2	2	1	$\frac{12}{21} = 0.571$
1	4	0	$\frac{8}{21} = 0.381$

These should match the proportions found by repeating shufflings of the deck of 10 allele cards.

Permutation Test for D3S1358

For an STR locus, where $\{n_g\}$ are the genotype counts and $n = \sum_g n_g$ is the sample size, and $\{n_a\}$ are the alleles counts with $2n = \sum_a n_a$, the exact test statistic is

$$\Pr(\{n_g\}|\{n_a\}, \text{HWE}) = \frac{n! 2^H \prod_a n_a!}{\prod_g n_g! (2n)!}$$

where H is the count of heterozygotes.

This probability for the African American genotypic counts at D3S1358 is 0.6163×10^{-13}, which is a very small number. But it is not unusually small if HWE holds: a proportion 0.81 of 1000 permutations have an even smaller probability. We do not reject the HWE hypothesis in this case.

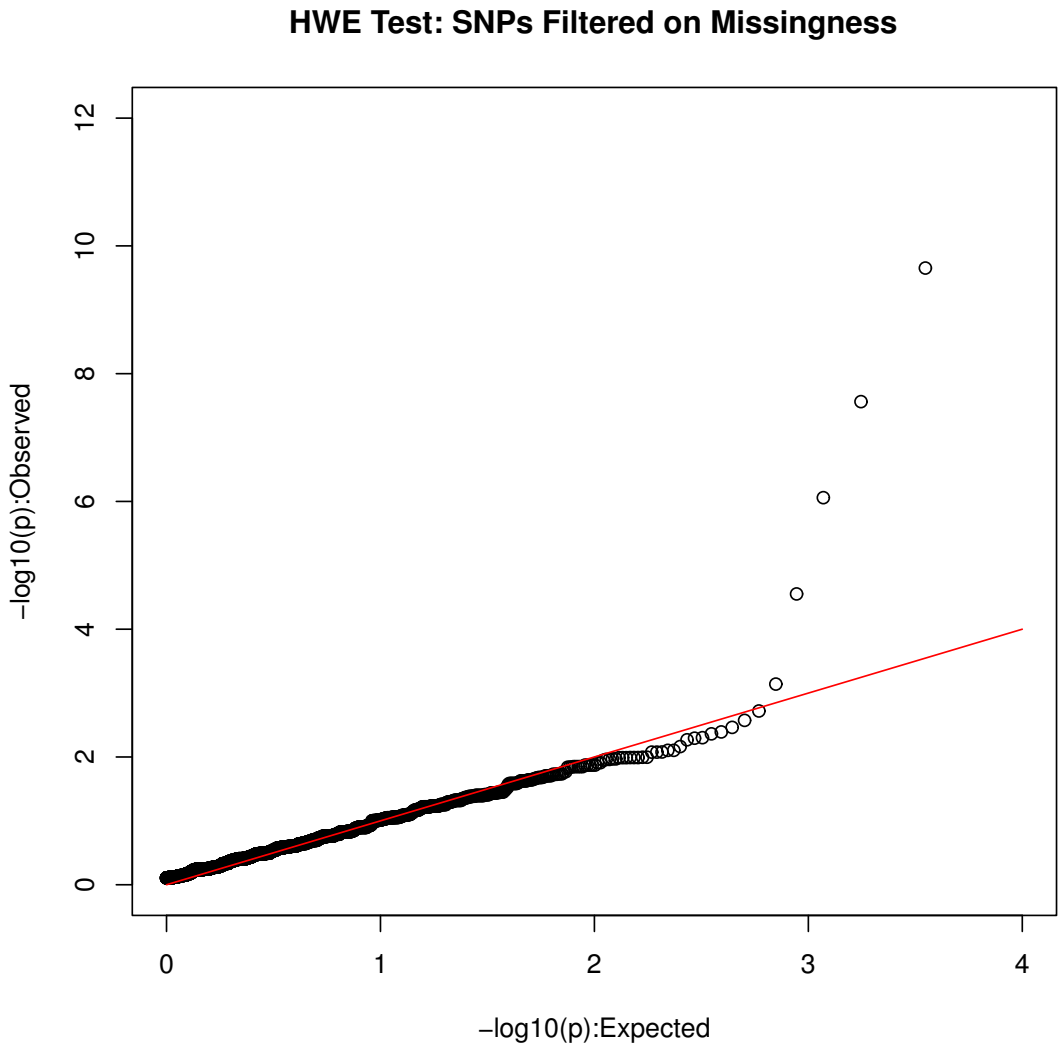
Linkage Disequilibrium

This term is generally reserved for association between pairs of alleles – one at each of two loci. In the present context, it may simply mean some lack of independence of profile or match probabilities at different loci.

Unlinked pairs of loci are expected to be almost independent, and two-locus tests are generally not significant.

However, if two profiles match at several loci this may be because they are from the same, or related, people and so are likely to match at additional loci. We do not have good tests of linkage disequilibrium across multiple loci.

QQ Plots for Large SNP Studies



QQ Plots for NIST 1036 Hispanics

Locus	HWE	EXP
F13B	0.0397	0.0172
TH01	0.0916	0.0517
F13A01	0.1050	0.0862
PentaD	0.1353	0.1207
PentaC	0.1369	0.1552
D22S1045	0.2359	0.1897
TPOX	0.2778	0.2241
D8S1179	0.3322	0.2586
D16S539	0.4475	0.2931
LPL	0.4516	0.3276
D1S1656	0.4806	0.3621
D19S433	0.5275	0.3966
D12S391	0.5516	0.4310
D21S1338	0.5547	0.4655
PentaE	0.6644	0.5000
FESFPS	0.6644	0.5345
D6S1043	0.6691	0.5690
D5S818	0.6691	0.6034
D2S441	0.7247	0.6379
D7S820	0.7269	0.6724
CSF1PO	0.7781	0.7069
FGA	0.7919	0.7414
D21S11	0.8219	0.7759
D10S1248	0.8238	0.8103
SE33	0.8784	0.8448
DSS1358	0.8950	0.8793
D18S51	0.9116	0.9138
vWA	0.9116	0.9483
vWA	0.9872	0.9828

QQ Plots for NIST 1036 All

Locus	HWE	EXP
CSF1PO	0.0000	0.0172
D10S1248	0.0000	0.0517
D12S391	0.0000	0.0862
D13S317	0.0006	0.1207
D16S539	0.0047	0.1552
D18S51	0.0075	0.1897
D19S433	0.0184	0.2241
D1S1656	0.0444	0.2586
D21S11	0.0656	0.2931
D22S1045	0.0709	0.3276
D21S1338	0.0750	0.3621
D2S441	0.1281	0.3966
DSS1358	0.1556	0.4310
D5S818	0.1863	0.4655
D6S1043	0.2078	0.5000
D7S820	0.2200	0.5345
D8S1179	0.2675	0.5690
F13A01	0.2913	0.6034
F13B	0.3272	0.6379
FESFPS	0.3813	0.6724
FGA	0.4241	0.7069
LPL	0.4431	0.7414
PentaC	0.4588	0.7759
PentaD	0.5066	0.8103
PentaE	0.5144	0.8448
SE33	0.5163	0.8793
TH01	0.6678	0.9138
TPOX	0.7522	0.9483
vWA	0.7644	0.9828

Hardy-Weinberg Equilibrium

HWE is a basic law in population genetics, described in 1908 by English mathematician Hardy and German physician Weinberg.

If A and B are two alleles for a gene, and if they have population proportions p_A and p_B then the population proportion of the three genotypes AA, AB and BB are p_A^2, $2p_Ap_B$, and p_B^2.

This law assumes infinitely large populations, mating at random, with no migration or mutation or natural selection. It is not 'true' but it fits very well to data from human populations.

Response to Lander

Is HWE expected? Human populations are generally found to be in HWE.

Does a departure from HWE indicate population structure? Although population structure can lead to departures from HWE, so can several other factors.

Why were departures detected from HWE? It is often the case that electrophoretic detection of RFLP alleles results in heterozygotes being classified as homozygotes. "The Lifecodes database of three VNTR loci used for forensics was used to show that the claimed excess of homozygotes is not necessarily real because many heterozygotes with similar allele sizes are misclassified as homozygotes. A simple test of H-W that takes such misclassifications into account was developed to test for an overall excess or dearth of heterozygotes in the sample (the complement of homozygote dearth or excess). The application of this test to the Lifecodes database revealed that there was no consistent evidence of violation of H-W for the Caucasian, black, or Hispanic populations."

Devlin B, et al. 1990. Science 249:1416-1420.

Hardy-Weinberg Testing

The Lifecodes data were for VNTR markers, where the repeat units were long and there were many repeat units. It was often difficult to distinguish alleles with different, but not very different, numbers of repeat units. Also, sometimes short alleles ran off the end of the gel. VNTR data are not used today.

To illustrate HWE, we will look at some FBI data for a microsatellite marker with 10 alleles. Such a marker has 10 homozygote types and 45 heterozygote types, and it is quite likely that several genotypes will not be seen in a sample of a few hundred individuals even though all the alleles are seen.

For example, consider these genotype counts for the D3S1358 marker published by the FBI (Budowle and Moretti, 1999) for an African-American sample:

http://www.fbi.gov/hq/lab/fsc/backissu/july1999/budowle.htm

Hardy-Weinberg Testing

The 55 genotype counts, arranged with allele type headings for rows and columns:

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

Hardy-Weinberg Testing

31 of the 55 genotypes did not appear in the sample of 210 individuals. How often would one of those types occur in the population? HWE lets us calculate that from the observed allele frequencies.

The allele counts in the data 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

Hardy-Weinberg Testing

In these data there are 210 genotypes: 45 homozygotes and 165 heterozygotes. The number of homozygotes expected under HWE is the sample size times the sum of squares of allele frequencies:

$$\begin{aligned} & 210 \times \left[\left(\frac{2}{420}\right)^2 + \left(\frac{1}{420}\right)^2 + \left(\frac{5}{420}\right)^2 + \left(\frac{51}{420}\right)^2 + \left(\frac{122}{420}\right)^2 \right. \\ & \left. + \left(\frac{129}{420}\right)^2 + \left(\frac{84}{420}\right)^2 + \left(\frac{23}{420}\right)^2 + \left(\frac{2}{420}\right)^2 + \left(\frac{1}{420}\right)^2 \right] \\ = \; & 210 \times \frac{41746}{176400} \\ = \; & 49.7 \end{aligned}$$

Hardy-Weinberg Testing

There is good agreement between observed and expected homozygote counts, as expected. A formal goodness-of-fit statistical test is conducted as

	Observed	Expected	$\frac{(o-e)^2}{e}$
Homozygotes	45	49.7	0.44
Heterozygotes	165	160.3	0.14
Total	210	210.0	0.58

The test statistic $X^2 = 0.58$ is from a chi-square distribution with 1 degree of freedom under the null hypothesis of HWE. It would need to be at least 3.84 to declare significance at the 5% level. No basis for rejecting HWE in this case.

There are several other approaches to testing for HWE.

Predicting Genotype Counts

Suppose a crime stain had the D3S1358 type 12,15 (heterozygous for alleles 12 and 15). None were observed in the FBI data, but alleles 12 and 15 were seen with sample proportions 1/420 and 122/420.

The HWE population proportion is $2 \times \frac{1}{420} \times \frac{122}{420} = 0.0014$. There is only a small probability a random person would have that type, so if a suspect has that type then the evidence seems quite strong.

Predicting Profile Counts

The power of DNA profiling comes from using many loci: the FBI developed a 13-marker (now 20) "CODIS" panel. To predict a 13-marker profile probability it is usual to multiply over markers. It is easy to test for HWE, independence of the two alleles, at one marker but not possible to test for linkage equilibrium, independence for all 26 alleles in a profile.

Indirect arguments support good consistency with independence.

FBI Database (94 Profiles) Matching Counts

Match. loci		Number of Partially Matching Loci													
		0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	O	0	3	18	92	249	624	1077	1363	1116	849	379	112	25	4
	E	0	2	19	90	293	672	1129	1403	1290	868	415	134	26	2
1	O	0	12	48	203	574	1133	1516	1596	1206	602	193	43	3	
	E	0	7	50	212	600	1192	1704	1768	1320	692	242	51	5	
2	O	0	7	61	203	539	836	942	807	471	187	35	2		
	E	1	9	56	210	514	871	1040	877	511	196	45	5		
3	O	0	6	33	124	215	320	259	196	92	16	1			
	E	1	7	36	116	243	344	334	220	94	23	3			
4	O	1	5	17	29	54	82	67	16	6	0				
	E	0	3	15	40	70	81	61	29	8	1				
5	O	0	1	2	6	12	14	6	5	0					
	E	0	1	4	9	13	11	6	2	0					
6	O	0	1	0	2	2	0	0	0						
	E	0	0	1	1	1	1	0	0						

What are match probabilities?

Identifiler [Applied Biosystems] 15 STR Loci Kit

Information is tied together with multiplex PCR and data analysis

D8S1179	{15,16}
D21S11	{29,29}
D7S820	{9,11}
CSF1PO	{10,11}
D3S1358	{16,17}
TH01	{6,7}
D13S317	{8,12}
D16S539	{10,11}
D2S1338	{19,19}
D19S433	{14,16}
VWA	{15,17}
TPOX	{8,12}
D18S51	{11,15}
Amel	{X,Y}
D5S818	{9,11}
FGA	{19,22}

Multiplying the frequency of each genotype at each locus gives us the Random Match Probability (RMP) of 1.25×10^{-15} for **unrelated individuals**

*The chance of an **unrelated individual** having this exact same profile is **1 in 800 trillion***

This test contains the 13 FBI core loci

[Vallone: https://www.nist.gov/document-7351]

Will match probabilities keep decreasing?

STR panel	Match probability
13-locus CODIS	2.34×10^{-15}
15-locus Identifiler	5.93×10^{-18}
20-locus CODIS	9.54×10^{-25}
24-locus FBI core	6.28×10^{-30}

[Ge et al, Investigative Genetics 3:1-14, 2012]

Will match probabilities keep decreasing?

How do these match probabilities address the observation of Donnelly:

> "after the observation of matches at some loci, it is relatively much more likely that the individuals involved are related (precisely because matches between unrelated individuals are unusual) in which case matches observed at subsequent loci will be less surprising. That is, knowledge of matches at some loci will increase the chances of matches at subsequent loci, in contrast to the independence assumption."

[Donnelly, Heredity 75:26-64. 1995]

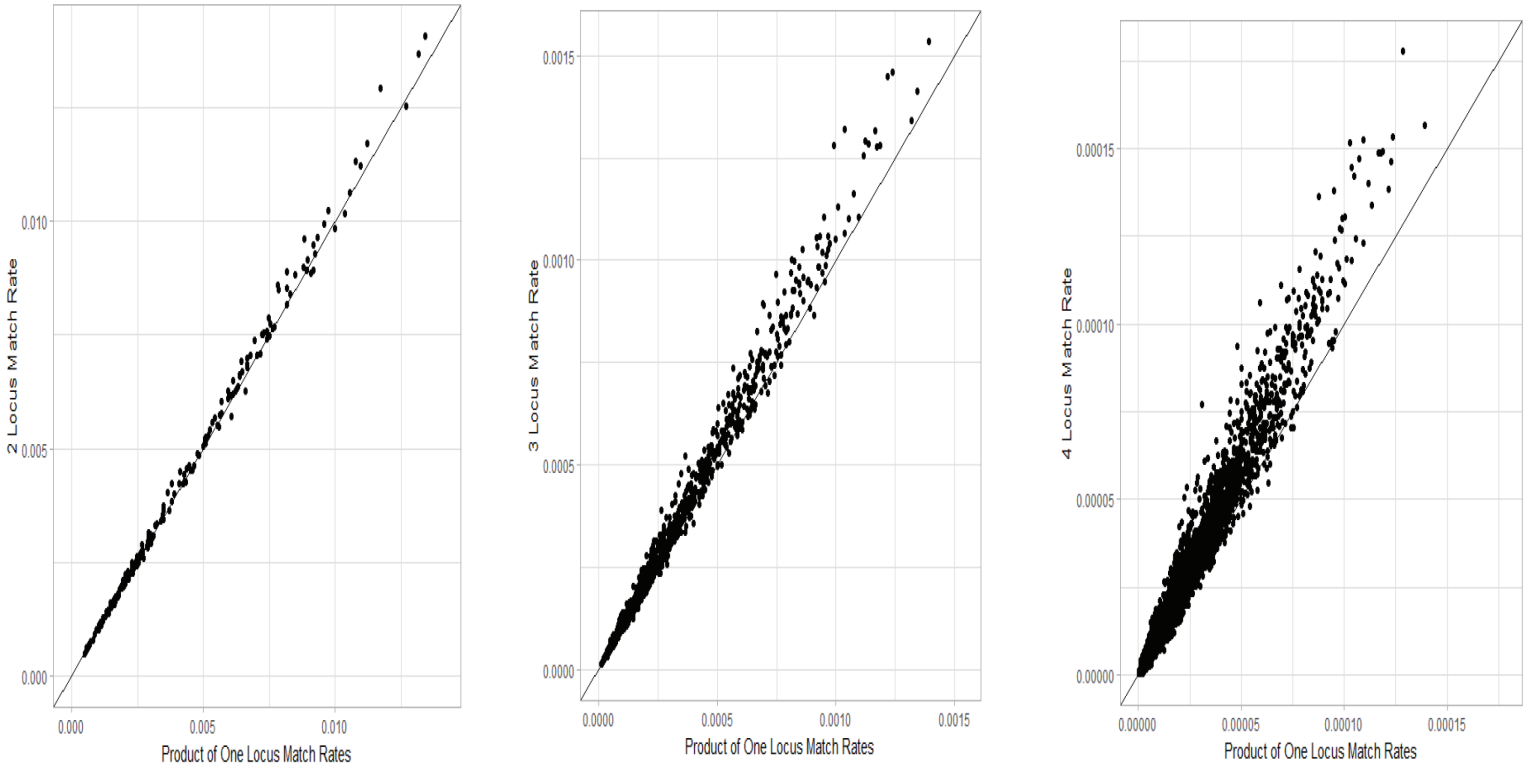
Are match probabilities independent over loci?

Is the problem that we keep on multiplying match probabilities over loci under the assumption they are independent? Can we even test that assumption for 10 or more loci?

Or is the standard "random match probability" not the appropriate statistic to be reporting in casework? Is it actually appropriate to report statements such as

> The approximate incidence of this profile is 1 in 810 quintillion Caucasians, 1 in 4.9 sextillion African Americans and 1 in 410 quadrillion Hispanics.

Empirical dependencies: 2849 20-locus profiles



The product over loci under-estimates the actual match proportions to an extent that increases with more loci.

Edward Zhao, unpublished

Match Probabilities

The match probability is usually estimated using allele frequencies from a database representing some broad class of people, such as "Caucasian" or "African American" or "Hispanic."

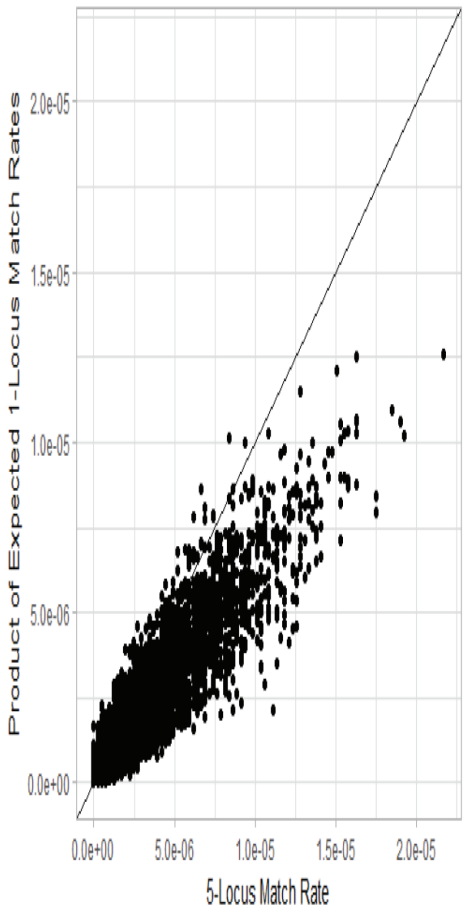
The population relevant for a particular crime may be a narrower class of people. There is population structure, quantified by the parameter θ. If p are the allele frequencies in the database, the match probabilities are estimated as

$$\begin{aligned}
\Pr(AA|AA) &= \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)} \\
\Pr(AB|AB) &= \frac{2[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)}
\end{aligned}$$

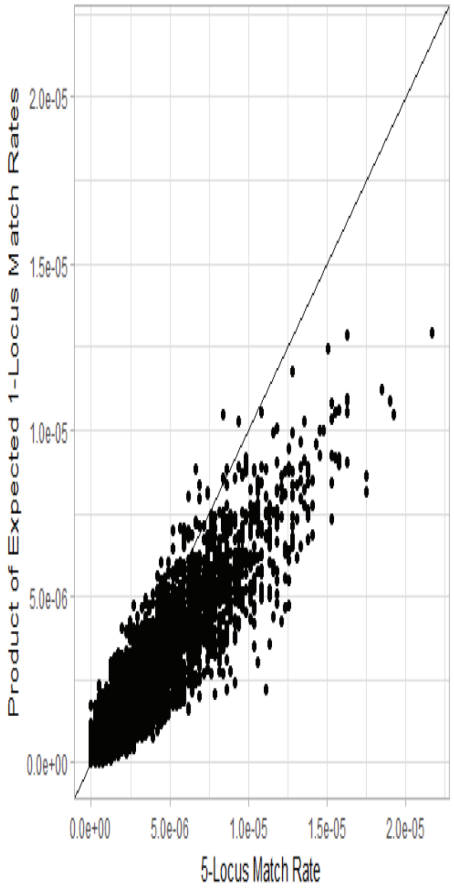
Can these be multiplied over loci?

2849 US profiles

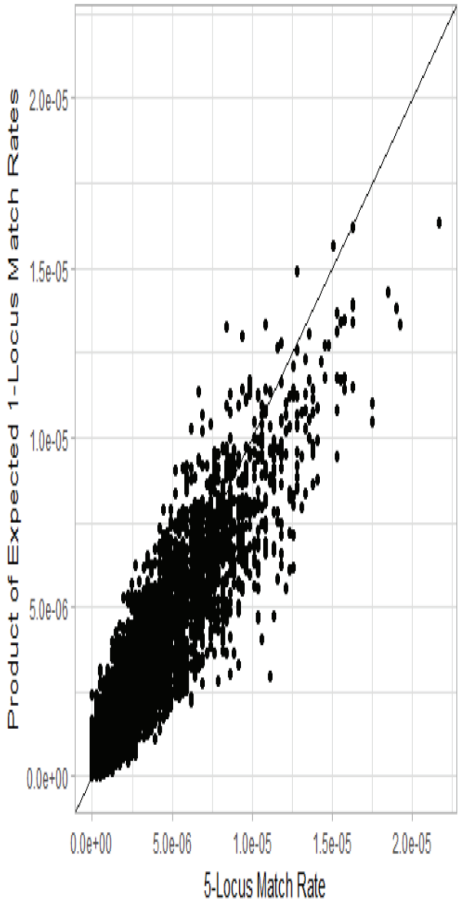
$\theta = 0$



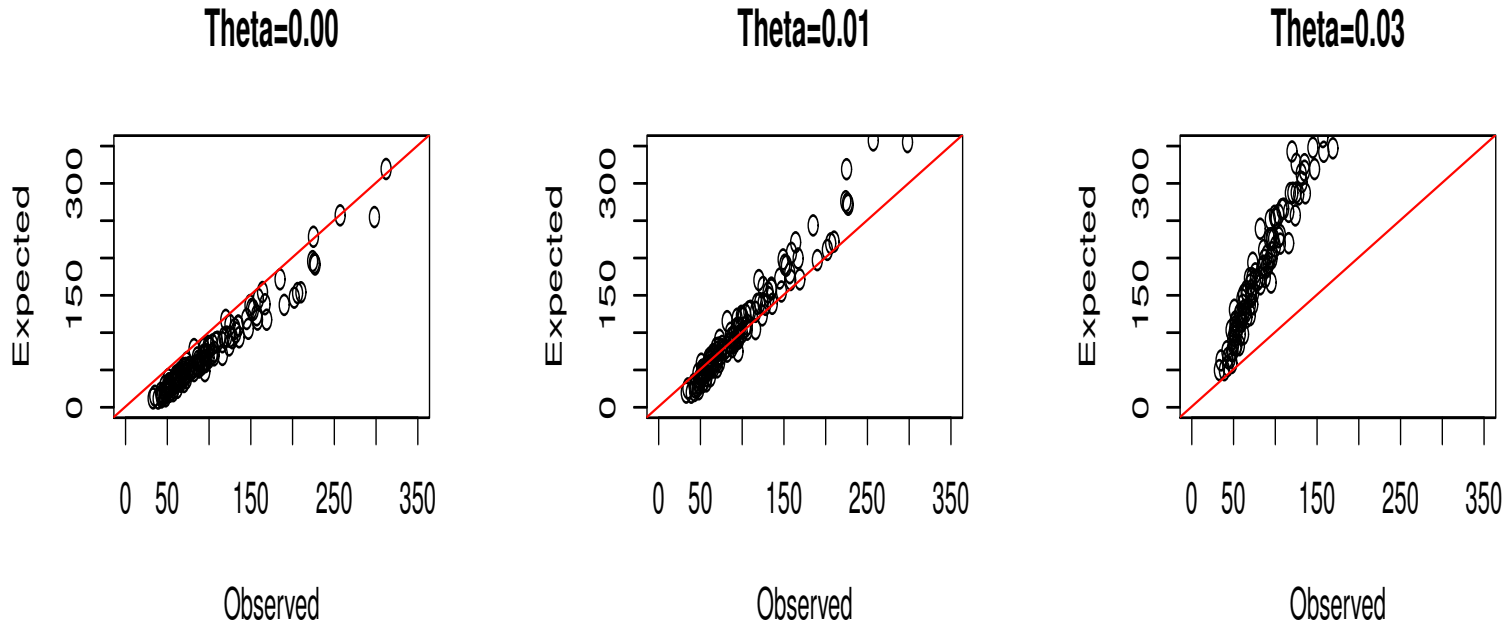
$\theta = 0.001$



$\theta = 0.01$



15,000 Australian Profiles



Numbers of five-locus matches among nine-locus profiles.

Weir BS. 2004. Journal of Forensic Sciences 49:1009-1014

Conclusions

- Product of profile probabilities decreases at the same rate as number of loci increases.
- Match probabilities are not profile probabilities.
- Match probabilities decrease more slowly as number of loci increases.
- "Theta correction" may accommodate multi-locus dependencies.
- Empirical studies need much larger databases.