703 Notes John Buckleton 2009

Allele	Technically this refers to the different forms of a gene. However in forensic DNA profiling it is misused to refer to the different forms of the intron, which, technically, is not a gene.
Autosomes	Any pair of chromosomes other than the XY pair
Bayes theorem	A mathematical theorem developed by the Reverend Bayes stating that the posterior odds are equal to the prior odds multiplied by the likelihood ratio
Billion	1,000,000,000
Population Bottleneck	A **population bottleneck** (or **genetic bottleneck**) is an evolutionary event in which a significant percentage of a population or species is killed or otherwise prevented from reproducing, and the population is reduced by 50% or more, often by several orders of magnitude.
Chromosome	A physical structure of the nucleus that contains the DNA sequence. From the Latin for a coloured body from their affinity to take up dye.
Diploid	Describes an organism or cell with two copies of each chromosome.
Gene flow	Gene flow is the exchange of genes between populations, which are usually of the same species. It may occur with or without the physical movement of individuals
Genetic Drift	Genetic drift is the accumulation of purely random changes in relative abundance of allele frequencies in a population
Founder effect	The establishing a new population by a small number of individuals
Gamete	The reproductive cells: an egg or a sperm. These are haploid.
Gene duplication	
Gene translocation	
Chromosomal rearrangement	
Gonosomes	This refers to the XY chromosome pair
Homozygote	The genotype at this locus has two copies of the same allele
Haploid	An organism or cell with a single copy of each chromosome.
Hardy Weinberg Equilibrium	An assumption of independence at one locus.
Heterozygote	The genotype at this locus has two different alleles
Linkage equilibrium	An assumption of independence between loci.
Loci/Locus	A position on the genome (loci is the plural)
Mendelian inheritance	Inheritance that follows Mendel's two laws.
Mitochondria	An organelle in eukaryotes associated with the production of ATP.
Mitochondrial DNA	The DNA present as small circular molecules in the mitochondria
Mitotype	The genotype of the mitochondrial DNA.
Mosaic trisomy	
mtDNA	Mitochondrial DNA
Partial trisomy	

Paternal inheritance	Inheritance from the father
paternity index	A term used in paternity testing for the likelihood ratio
Posterior odds	Usually referring to Bayes theorem
Prior odds	Usually referring to Bayes theorem
Probability of paternity	A term used in paternity testing for the posterior probability of paternity given prior odds of 1.
Product rule	Product rule. The assumption of hardy-Weinberg and linkage equilibrium together is the product rule
Punnett square	A method for assigning the probabilities of children conditional on their parents' genotypes
r or R_c	The recombination fraction
Trisomy	The situation where an individual has three copies of a chromosome rather than the usual pair.
θ	the probability that two alleles taken from two individuals of the same sub-population are identical by descent (ibd)
Wahlund effect	Wahlund effect refers to reduction of heterozygosity in a population caused by subpopulation structure

LAWS OF PROBABILITY

Nomenclature

Event, say, A.

Example: A: the next card is an ace.
Hp: The suspect is the donor of the DNA

Events may be true or false, they may be in the past, present or future

Pr(A) is shorthand for: The probability that event A is true.

$\Pr(\bar{A})$ is shorthand for: The probability that A is false

$$\Pr(\bar{A}) = 1 - \Pr(A)$$

Two events, A and B can be written A&B, or A,B

1st law. $0 \le \Pr(A) \le 1$

2nd law. For two events A and B the probability that A or B is true

Pr(A or B) = Pr(A)+Pr(B)-Pr(A,B)

Mutually exclusive *if two or more events are mutually exclusive then the occurrence of one of them means that none of the others can occur*

If A and B are mutually exclusive then

Pr(A or B) = Pr(A)+Pr(B)

Example. What is the probability of throwing a 2 or a 4 with a fair six sided dice.

Ask are they mutually exclusive? Can you get a 2 and a 4 at the same time? No. So they are mutually exclusive. Let

2: be the event that the throw is a 2 AND
4: be the event that the throw is a 4 then

Pr(2 or 4) = 1/6 + 1/6 = 2/6 = 1/3

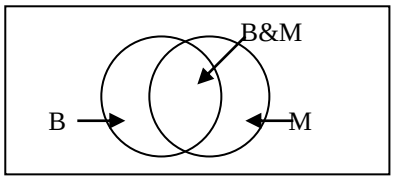
3rd law: before proceeding to the third law we will need to understand conditional probability

Conditional probability.

There are available several definitions for conditional probability. One proceeds from the third law of probability

$$\Pr(a \mid b) = \frac{\Pr(a,b)}{\Pr(b)} \quad \ldots\ldots 3^{\text{rd}} \text{ law of probability}$$

which can be interpreted quite well in set theory. For instance evaluating Pr(*a*|*b*) involves enumerating the set of outcomes where event *b* is true and seeing in what fraction of these event *a* is also true.



Example: In a certain office there are 10 men. Three men have beards (event *B*) and moustaches (event *M*). A further two have moustaches only. Say we were interested in Pr(*B*|*M*) we find the set of men where M is true: this has 5 members. Of these 3 have beards. Hence Pr(*B*|*M*)=3/5.

If we were interested in Pr(*M*|*B*) we find the set of men where *B* is true: this is 3 men. Of these all 3 have moustaches. Hence Pr(*M*|*B*) = 3/3 = 1.[1]

Example:

Ethnic group and age data from the 1996 census

	0-14	15-64	65+	Total
Caucasian	13.8%	47.5%	10.4%	71.7%
Maori	5.5%	8.6%	0.4%	14.5%
Pacific Islander	1.7%	2.9%	0.2%	4.8%
Others	2.0%	6.3%	0.7%	9.0%
Total		23.0%	65.3%	11.7% 100.0%

Pr[Maori] = 0.145 Pr[0 - 14] = 0.230
Pr[Maori and 0 - 14] = 0.055
Pr[Maori | 0 - 14] = 0.055/0.230
Pr[0 - 14 | Maori] = 0.055/0.145

3rd law: Pr(A,B) = Pr(A|B)Pr(B) = Pr(B|A)Pr(A)

Variants that may be useful

Pr(A,B,C) = Pr (A|B,C)Pr(B,C)

[1] In this simple example we are making an assumption that each of the men is equally likely to be observed. This assumption may not be true in more general examples, but the principle behind the definition of the conditional probability remains valid.

$Pr(A,B,C,D) = Pr(A|B,C,D)Pr(B,C,D)$
$Pr(A,B,C,D,E) = Pr(A|B,C,D,E)Pr(B,C,D,E)$

$Pr(A,B,C|I) = Pr(A|B,C,I)Pr(B,C|I)$

INDEPENDENCE

Two events are independent if the truth of one does not affect the uncertainty of the other.

If two events are independent then

$Pr(A,B) = Pr(A)Pr(B)$

How to check for independence: Check if $Pr(A,B) = Pr(A)Pr(B)$

You may also do this using logic. For example the roll of two dice. The number on one cannot affect the number on the other.

Example: P v Collins

Victim has purse stolen by Caucasian woman with a blonde ponytail who gets into a yellow car with a bearded negro man.

An instructor or mathematics assumes the following numbers and independence which gives 1 in 12,000,000:

A: Yellow automobile	0.10
B: Man with moustache	0.25
C: Girl with a pony tail	0.10
D: Girl with blonde hair	0.33
E: African American male with beard	0.10
F: Interracial couple	0.001

Are the events independent? For example if you know E does that change the uncertainty in B. If you know E and D does that change the uncertainty in F. Can we do this one properly?

Take the most dependent event out first. I know this from experience you will have to work through it logically.
$Pr(A,B,C,D,E,F) = Pr(F|A,B,C,D,E)Pr(A,B,C,D,E)$

$Pr(F|A,B,C,D,E)\approx 1$ since if I knew the woman was blonde and the man African American then I can be reasonably sure that the couple are interracial.

$Pr(A,B,C,D,E)= Pr(B|A,C,D,E)Pr(A,C,D,E)$

If an event has no effect on the event in front of the conditioning bar you can remove it from behind the bar.

$Pr(B|A,C,D,E)=Pr(B| E)$

Since I assume that car and the girl's hair colour and pony tail have nothing to do with whether he has a moustache. But having a beard does.

Pr(A,C,D,E) =Pr(A)Pr(C)Pr(D)Pr(E)

I think the last four events are all independent or nearly so.

Pr(A,B,C,D,E,F) = 1 x Pr(B|D)Pr(A)Pr(C)Pr(D)Pr(E)

Example from the 2007 exam
1. Please answer all parts of part (a) and part (b).
(a) In People v Collins an eyewitness saw a blonde female with a ponytail rob an elderly woman and get into a partly yellow car driven by a male African-American with a beard and moustache. Later a couple was apprehended who nearly fitted this description. An instructor of mathematics famously interpreted the case by assuming these probabilities:

Event		probability
A	Partly yellow automobile	0.10
B	Man with moustache	0.25
C	Girl with pony tail	0.10
D	Girl with blond hair	0.33
E	African-American man with beard	0.10
F	Interracial Couple in a car	0.001

(i) When he multiplied these probabilities what did he assume? (2 marks)

Independence

(ii) Using the events A,B,C,D,E,F please show how to break down Pr(A,B,C,D,E,F) using the third law of probability. (2 marks)

Any order of decomposition accepted. Such as
$$\Pr(A,B,C,D,E,F) = \Pr(A \mid B,C,D,E,F)\Pr(B \mid C,D,E,F)\Pr(C \mid D,E,F)\Pr(D \mid E,F)\Pr(E \mid F)\Pr(F)$$

In a survey of 100 African males the following results were obtained:

		Moustache	
		No	Yes
Beard	No	72	12
	Yes	1	15

(iii)Please estimate the probability that an African-American man has a beard.
(1 mark)

$$\frac{16}{100}$$

(iv)What type of probability is this? (1 mark)

I have not taught subjective and objective probabilities in 2009

(v) Please estimate Pr(B|E) from this data. (2 marks)

$\frac{15}{16}$

(vi) Why is this different from what the instructor assumed? (2 marks)

Because beards and moustaches are dependent.

Example from the 2008 exam

(b) A crime is observed in Remuera, Auckland, where a balding man with estimated age 55 - 65 years old is seen to drive a black Porche into a small child playing by the road. He drives away without ascertaining if injury has happened. At a later stage a balding man, Mr Smith, aged 53 who drives a black Porche is found rapidly cleaning parts of his car.

An instructor of mathematics reports that he sat on Symonds Street and counted one black Porche in 1,000 cars. He then observed the people and saw only 12 men in 1000 that he might say were 55 - 65 years old. He then watched another 1000 men and 3 were balding.

He estimated that the chance of a balding man with estimated age 55 – 65 years who drives a black Porche is

$$\frac{1}{1,000}\times\frac{12}{1,000}\times\frac{3}{1,000}=\frac{36}{1,000,000,000}$$ or 36 in a billion.

He states in court that the chance that it is not Mr Smith who hit the child is 36 in a billion.

E_1: estimated age of offender is 55 years
E_2: offender is described as balding
E_3: offender drives a black Porche

And the hypotheses

Hp: Mr Smith is the offender
Hd: A random man is the offender

(i) Please use the third law to decompose (break down into parts) the likelihood ratio. You are not expected to produce numerical results. (4 marks)

$$\frac{\Pr(E1,E2,E3\mid Hp)}{\Pr(E1,E2,E3\mid Hd)}=\frac{\Pr(E1\mid E2,E3,Hp)}{\Pr(E1\mid E2,E3,Hd)}\times\frac{\Pr(E2\mid E3,Hp)}{\Pr(E2\mid E3,Hd)}\times\frac{\Pr(E3\mid Hp)}{\Pr(E3\mid Hd)}$$

Any order accepted but this is my favourite

(ii) Please discuss which factors may be expected to be independent or dependant (4 marks)

I expect age and balding to be dependent. I think age and Porche and even balding and Porche might also be.

(iii) Has the instructor of mathematics tried to calculate the likelihood ratio, the numerator of a likelihood or the denominator of a likelihood ratio? Which factor(s) has he not assessed at all? (2 marks)

I think it is an attempt at the denominator of the LR with no effort at the numerator or at dependencies.

(iv) Please critique the work of the instructor of mathematics. (4 marks)

He has assumed independence without any proof and probably wrongly.

He has not assessed the numerator of the LR.
He has not considered the appropriateness of his survey to the crime.

Example from the 2009 exam

Q3b
Video surveillance of a dangerous street corner shows a man with a beard and moustache assault and kill a pedestrian and steal his wallet. The video is clear enough to state with certainty that the man has a beard and a moustache but nothing else can be told. A suspect Mr A is found who has a moustache and beard. Let

B: Be the event that the man has a beard
M: Be the event that the man has a moustache
Hp: Mr A is the man who assaulted and killed the pedestrian
Hd: A random man assaulted and killed the pedestrian

Survey data was taken in the area of the crime at a similar time. It is given below.

	Beard	No beard	Total
Moustache	9	5	
No Moustache	1	85	
Total			100

i)What is the probability of a beard Pr(B)? 0.10 1mark
ii)What is the probability of a moustache PrM)? 0.14 1mark
iii)What is the probability of a moustache given a beard Pr(M|B)? 0.90 1mark
Are beard and moustache independent? How could you check?

No. Compare Pr(M|B) with Pr(M) or many other ways. 1marks

Please evaluate the likelihood ratio $LR = \frac{\Pr(M, B \mid Hp)}{\Pr(M, B \mid Hd)}$? $LR = \frac{1}{0.09} = 11.1$

4marks

Please correctly phrase the answer.

The evidence is 11 times more likely if Mr A is the person who assaulted and killed the pedestrian than if a random man is the person who assaulted and killed the pedestrian.

4marks

ODDS AND PROBABILITIES

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

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

COINCIDENCE PROBABILITIES

The coincidence approach proceeds to offer evidence against a proposition by showing that the evidence is unlikely if this proposition is true. Hence it supports the alternative proposition. The less likely the evidence is under the proposition the more support that is given to the alternative.

This is called the coincidence probability approach because either the evidence came from, say, the suspect or a 'coincidence' has occurred.

There are many examples of evidence presented in this way:

- 'only 1% of glass would match the glass on the clothing by chance,'
- 'it is very unlikely to get this paint sequence match by chance alone,'
- 'approximately 1 in a million unrelated males would match the DNA at the scene by chance.'

We are led to believe that the event 'match by chance' is unlikely and hence the evidence supports the alternative. At this stage let us proceed by assuming that if the evidence is unlikely under a particular hypothesis then this supports the alternative.

This is strongly akin to formal hypothesis testing procedures in statistical theory. Formal hypothesis testing would proceed by setting up the hypothesis usually called the null, H_0. The probability of the evidence (or data) is calculated if H_0 is true. If this probability is small (say less than 5% or 1%) then the null is 'rejected.' The evidence is taken to support the alternative hypothesis, H_1 [305,579,612]

To set up a DNA case in this framework we could proceed:

Formulate the hypothesis, H_0: the DNA came from a male not related to the suspect.

We then calculate the probability of the evidence if this is true. We write the evidence as E, and in this context it will be something like:

E: The DNA at the scene is type α.

We assume that it is known that the suspect is also type α. We calculate the probability, Pr, of the evidence, E, if the null hypothesis H_0 is true, $\Pr(E \mid H_0)$. The vertical line, or conditioning sign, stands for the word 'if' or 'given'.

Assuming that about 1 in a million unrelated males would have type α we would assign $\Pr(E|H_0)$ as 1 in a million. Since this is a very small chance we would assume that this evidence suggests that H_0 is not true and hence is support for H_1. In this context we might define the alternative hypothesis as:

H_1: the DNA came from the suspect.

Hence in this case the evidence supports the hypothesis that the DNA came from the suspect. Later we are going to need to be a lot more careful about how we define hypotheses.

Hypothesis testing is a well-known and largely accepted statistical approach. The similarity between the coincidence approach and hypothesis testing is the former's greatest claim to prominence.

BAYESIAN APPROACH

Frustrations with the frequentist approach to forensic evidence have led many people to search for alternatives.[105,258] For many these frustrations stem from discussing multiple stains, multiple suspects, or from trying to combine different evidence types.[652,656] The foremost alternative is the logical approach (also called the Bayesian approach).[257,490,500,516,517,518] This approach has been implemented routinely in paternity cases since the 1930's.[255] It is however only in the latter stages of the 20th century that it made inroads into many other fields of forensic science. It now dominates the forensic literature, but not necessarily forensic practice, as the method of choice for interpreting forensic evidence.[6,8,170,171,173,214,335,585,659,663] Bär gives an elegant review.[47]

Let:

H_p: Be the hypothesis advanced by the prosecution.

H_d: Be a particular hypothesis suitable for the defence.

E: Represent the evidence.

I: Represent all the background evidence relevant to the case.

The laws of probability lead to

$$\frac{\Pr(H_p \mid E,I)}{\Pr(H_d \mid E,I)} = \frac{\Pr(E \mid H_p,I)}{\Pr(E \mid H_d,I)} \times \frac{\Pr(H_p \mid I)}{\Pr(H_d \mid I)} \quad \text{....equation 2.1.}$$

This theorem is known as Bayes' theorem.[53] A derivation appears in Box 2.1. This theorem follows directly from the laws of probability. It can therefore be accepted as a logical framework for interpreting evidence.

Box 2.1 A derivation of Bayes' theorem

The third law of probability states:

$$\Pr(a \text{ and } b \mid c) = \Pr(a,b \mid c) = \Pr(a \mid b,c)\Pr(b \mid c) = \Pr(b \mid a,c)\Pr(a \mid c)$$

rewriting this using H_p, H_d, E and I

$$\Pr(H_p,E \mid I) = \Pr(H_p \mid E,I)\Pr(E \mid I) = \Pr(E \mid H_p,I)\Pr(H_p \mid I)$$

and

$$\Pr(H_d,E \mid I) = \Pr(H_d \mid E,I)\Pr(E \mid I) = \Pr(E \mid H_d,I)\Pr(H_d \mid I)$$

hence

$$\frac{\Pr(H_p,E \mid I)}{\Pr(H_d,E \mid I)} = \frac{\Pr(H_p \mid E,I)\Pr(E \mid I)}{\Pr(H_d \mid E,I)\Pr(E \mid I)} = \frac{\Pr(E \mid H_p,I)\Pr(H_p \mid I)}{\Pr(E \mid H_d,I)\Pr(H_d \mid I)}$$

hence

$$\frac{\Pr(H_p \mid E,I)\Pr(E \mid I)}{\Pr(H_d \mid E,I)\Pr(E \mid I)} = \frac{\Pr(E \mid H_p,I)\Pr(H_p \mid I)}{\Pr(E \mid H_d,I)\Pr(H_d \mid I)}$$

cancelling $\Pr(E \mid I)$

$$\frac{\Pr(H_p \mid E,I)}{\Pr(H_d \mid E,I)} = \frac{\Pr(E \mid H_p,I)\Pr(H_p \mid I)}{\Pr(E \mid H_d,I)\Pr(H_d \mid I)} \quad eq.....2.1$$

Equation 2.1 is often given verbally as

$$posterior\ odds = likelihood\ ratio \times prior\ odds \quad \text{.....equation 2.2}$$

The prior odds are the odds on the hypotheses H_p before the DNA evidence. The posterior odds are these odds after the DNA evidence. The likelihood ratio informs us how to relate these two. This would seem to be a very worthwhile thing to do, that is, to relate the odds before consideration of the evidence to those after the evidence. It informs us how to update our opinion in a logical manner having heard the evidence.

The prior odds, $\frac{\Pr(H_p \mid I)}{\Pr(H_d \mid I)}$, represent the view on the prosecution and defence hypothesis before the DNA evidence is presented.[2] This view is something that is formed in the minds of the judge and jury. The information imparted to the jury is carefully restricted to those facts that are considered admissible and relevant. It is very unlikely that the prior odds are numerically expressed in the mind of the judge and jury and **there is no need that they should be numerical**.[662,663] Strictly it is not the business of the scientist to form a view on the 'prior odds' and most scientists would strictly avoid this (for a differing opinion see Meester and Sjerps[543] and the subsequent discussion[220]). These odds are based on the non-scientific evidence and it is the duty of judge and jury to assess this.[779,807]

The use of this approach typically reports only the likelihood ratio. By doing this the scientist reports the weight of the evidence without transgressing on those areas reserved for the judge and jury. This is the reason that the term 'the logical approach[3]' has been used to describe this method. It has also been described elsewhere as 'the likelihood ratio' approach. The term that is being avoided is 'the Bayesian approach' which is the term used in most papers on this subject including my own. This term is being avoided because, strictly, presenting a ratio of likelihoods does not necessarily imply the use of the Bayesian method. Most authors have intended the presentation of the likelihood ratio alone without necessarily implying that a discussion of Bayes' theorem and prior odds would follow in court. The intent was to present the scientific evidence in the context of a logical framework without necessarily presenting that framework.

To gain familiarity with equation 2.2 it is useful to consider a few results. What would happen if the likelihood ratio was 1? In this case the posterior odds are unchanged by the evidence. Another way of putting this is that the evidence is inconclusive.

What would happen if the likelihood ratio was greater that 1? In these cases the posterior odds would be greater than the prior odds. The evidence would have increased our belief in H_p relative to H_d. Another way of putting this is that the evidence supports H_p. The higher the likelihood ratio the greater the support for H_p.

If the likelihood ratio is less than 1 the posterior odds would be smaller than the prior odds. The evidence would have decreased our belief in H_p relative to H_d. Another way of putting this is that the evidence supports H_d. The lower the likelihood ratio the greater the support for H_d.

[2] My wording is wrongly implying an order to events such as the 'hearing of the DNA evidence'. In fact the evidence can be heard in any order. The mathematical treatment will give the same result regardless of the order in which the evidence is considered.[659]

[3] I first had this distinction explained to me by Dr Christophe Champod

Table 2.1: The prosecutor's nomogram. The prior and posterior probabilities associated with these odds are given next to the odds. Reproduced and amended from Riancho and Zarrabeitia[642]with kind permission of the authors and Springer-Verlag who retain ownership of the copyright.

Prior		Likelihood ratio	Posterior	
Probability	Odds		Odds	Probability
			100,000,000 to 1	99.999990%
0.001%	1 to 100,000		10,000,000 to 1	99.999989%
0.01%	1 to 10,000	10,000,000,000	1,000,000 to 1	99.9999%
		1,000,000,000		
0.1%	1 to 1,000	100,000,000	100,000 to 1	99.999%
		10,000,000		
1%	1 to 100	1,000,000	10,000 to 1	99.99%
		100,000		
9%	1 to 10	10,000	1,000 to 1	99.9%
		1,000		
50%	1 to 1	100	100 to 1	99%
		10		
91%	10 to 1	1	10 to 1	91%
99%	100 to 1		1 to 1	50%

It has been suggested that a nomogram may be useful to help explain the use of this formulation. This follows from a well-known nomogram in clinical medicine. Riancho and Zarrabeitia[642] suggest the diagram that has been modified and presented in Tables 2.1 and 2.2. These tables are used by choosing a prior odds and drawing a line through the centre of the *LR* value. The posterior odds may then be read directly. For example assume that the prior odds are about 1 to 100,000 (against) and the likelihood ratio is 10,000,000 then we read the posterior odds as 100 to 1 (on).

Table 2.2; The defendant's nomogram. Reproduced and amended from Riancho and Zarrabeitia[642] with kind permission of the authors and Springer-Verlag who retain ownership of the copyright.

Prior		Likelihood ratio	Posterior	
Probability	Odds		Odds	Probability
0.1%	1 to 1,000		100 to 1	99%
1%	1 to 100		10 to 1	91%
9%	1 to 10	10	1 to 1	50%
		1		
50%	1 to 1	1/10	1 to 10	9%
		1/100		
91%	10 to 1	1/1000	1 to 100	1%
		1/10,000		
99%	100 to 1	1/100,000	1 to 1,000	0.1%
		1/1,000,000		
99.9%	1,000 to 1	1/10,000,000	1 to 10,000	0.01%
		1/100,000,000		
99.99%	10,000 to 1	1/1,000,000,000	1 to 100,000	0.001%
			1 to 1,000,000	0.0001%

Table 2.3: A verbal scale

LR	**Verbal wording**	
1,000,000+	Extremely strong	Support for H_p
100,000	Very strong	
10,000	Strong	
1000	Moderately strong	
100	Moderate	
10	Limited	
1	Inconclusive	
0.1	Limited	Support for H_d
0.01	Moderate	
0.001	Moderately strong	
0.0001	Strong	
0.00001	Very strong	
0.000001	Extremely strong	

The likelihood ratio (*LR*) is a numerical scale. One point can be hinged to words without argument; a *LR* of 1 is inconclusive. Other words may be attached to this scale to give a subjective verbal impression of the weight of evidence.[12,94,174,263,264] This association of words with numbers is subjective and necessarily arbitrary. One such scale used extensively in the FSS is given in Table 2.3.

EXAMPLE OF THE BAYESIAN APPROACH

In the Bayesian approach the forensic scientist should concentrate only on the *LR*. The prior odds and hence the posterior odds are the province of the judge and jury.

Consider a stain at a scene that is type α. A suspect is found who is also type α. This is the evidence, E. It is composed from two parts

Gc: The crime stain is type α and
Gs: The suspect is type α.

We seek

$$LR = \frac{\Pr(E \mid Hp)}{\Pr(E \mid Hd)} = \frac{\Pr(Gc, Gs \mid Hp)}{\Pr(Gc, Gs \mid Hd)}$$

Using the 3rd law of probability we write

$$LR = \frac{\Pr(Gc \mid Gs, Hp)}{\Pr(Gc \mid Gs, Hd)} \times \frac{\Pr(Gs \mid Hp)}{\Pr(Gs \mid Hd)}$$

Whether the suspect is the donor of the stain should not affect his genotype so:

$$\Pr(Gs \mid Hp) = \Pr(Gs \mid Hd)$$

And hence

$$LR = \frac{\Pr(Gc \mid Gs, Hp)}{\Pr(Gc \mid Gs, Hd)}$$

But if the suspect is indeed the donor of the stain, *Hp*, then the crime stain must be the same as the suspect and hence type α. So

$$\Pr(Gc \mid Gs, Hp) = 1$$

If the suspect is not the donor of the stain then the stain at the scene must be type α because it has come from someone else who, by coincidence, is type α. Call this the frequency, *f*. Then

$$LR = \frac{1}{f}$$

Let us say that *f* is 1 in a billion then *LR* = one billion.

Correct phrasing, the evidence is a billion times more likely if the suspect is the donor of the stain than if a random person is the donor of the stain.

AN EXAMPLE FROM THE 2003 EXAM

a. A survey was undertaken of cars in a certain area. They were classified by colour as "White" or "Other" and by make as "Toyota" or "Other". Please complete this table by filling in the four empty cells. [2 marks]

	White	Other	Total
Toyota	26	74	100
Other	24	376	400
Total	50	450	500

There are four squares. ½ mark for each correct.

b. Take the table as representing the population of cars in some area. Please calculate Pr[Toyota], Pr[White], Pr[Toyota|White], and Pr[White|Toyota]. [2 marks]

Pr(Toyota)=100/500=0.2
Pr(White) = 50/500=0.1
Pr(Toyota|White)=26/50=0.52
Pr(White|Toyota)=26/100=0.26

There are four answers. Answer accepted as a fraction or a decimal. ½ mark for each correct.

c. A witness sees a crime in the area surveyed above and states that the get away vehicle was white. What are the posterior odds that it was a Toyota given the witness report that it was white? [3 marks]

Prior odds = 100:400

Posterior odds = 26:24=1.083. 3 marks for correct answer as 26:24 or 1.083. part marks for showing understanding.

Example from the 2004 exam

2. Please answer all parts.

a) A standard pack of 52 cards has been modified. The A♠ has been changed to the A♥ so now there are two A♥, no A♠, and one each of the other 50 cards. A card is drawn at random from this pack:

i) Are the events drawing a ♥ and drawing an ace, A, independent? Please show your working to justify your answer. (4 marks)

No. Pr(♥) = 14/52 Pr(A) = 4/52 Pr(♥&A) = 2/52 ≠ Pr(♥) x Pr(A)

ii) What is the probability that the drawn card is an A GIVEN that it is a ♥? (2 marks)
2/14

b) 90% of the taxis in a town are green and the rest are blue. According to an eyewitness, the perpetrator of a "hit and run" traffic offence was driving a blue taxi. This eyewitness

testifies honestly but he may have made a mistake about the colour of the taxi. Under these circumstances associated with this offence I estimate that the eyewitness may mistake a blue taxi for a green one and vice versa about 10% of the time.

i) What type of probability is the 10% estimate? (1 mark)

Not taught in 2009

ii) What are the prior odds that the perpetrator drove a blue taxi? (1 mark)

1:9 or 9 to 1 against

iii) What is the probability that the eyewitness would state that the taxi was blue GIVEN that it is in fact blue? (1 mark)

0.9

iv) What is the probability that the eyewitness would state that the taxi was blue GIVEN that it is in fact green? (1 mark)

0.1

v) What are the posterior odds on the taxi being blue GIVEN the eyewitness report that it was blue? (1 mark)

$\frac{1}{9}\times\frac{0.9}{0.1}=1$ or 1:1

vi) What is the probability that the perpetrator drove a blue taxi GIVEN the eyewitness report? (1 mark)

0.5

THE FALLACY OF THE TRANSPOSED CONDITIONAL

The fallacy of the transposed conditional comes from confusing the probability of the evidence given a specific hypothesis with the probability of the hypothesis itself. In the terms given above this would be confusing $\Pr(E \mid H_p)$ with $\Pr(H_p)$, $\Pr(H_p \mid E)$, or $\Pr(H_p \mid E, I)$.

Following a publication by Evett[256] we introduce the subject by asking

"*What is the probability of having four legs IF you are an elephant?*"

Let us write this as Pr(4|E) and we assign it a high value, say, 0.999.

Next we consider "what is the probability of being an elephant IF you have four legs?" Write this as Pr(E|4) and note that it is a very different probability and not likely to be equal to 0.999. This example seems very easy to understand both verbally and in the symbolic language of probability. But this fallacy seems to be quite tricky to avoid in court.

Imagine that we have testified in court along the lines of one of the statements given below:

- The probability of obtaining this profile from an unrelated male member of the New Zealand population is 1 in 3 billion.
- The frequency of this profile amongst members of the population of New Zealand unrelated to Mr Smith is 1 in 3 billion.
- This profile is 3 billion times more likely if it came from Mr Smith than if it came from an unrelated male member of the New Zealand population.

The first two are frequentist statements and the last is a statement of the likelihood ratio.

Let us work with the first. We are quite likely in court to face a question along the lines:

"In lay terms do you mean that the probability that this blood came from someone else is 1 in 3 billion?"

This is the fallacy of the transposed conditional. It has led to appeals and retrials. It appears to be very natural to make this transposition however incorrect. Every newspaper report of a trial that I have read is transposed and I suspect that many jurors and indeed judges make it.

How can a scientist who is testifying avoid this error?

The answer involves training and thinking on one's feet. But I report here Stella's Spotting Trick (named after Stella McCrossan) and Ian's coping trick (named after Ian Evett).

Stella's Spotting Trick: The key that Stella taught was to ask oneself whether the statement given is a question about the evidence or hypothesis. Probabilistic statements about the hypothesis will be transpositions. Those about the evidence are likely to be correct. The moment that you notice the statement does NOT contain an IF or a GIVEN you should be cautious. Consider the sentence given above: "In lay terms do you mean that the probability that this blood came from someone else is 1 in a billion?" Is this a statement about a proposition or the evidence? The proposition here is that the blood came from someone else. And indeed the statement is a question about the probability of the proposition. Hence it is a transposition.

Ian's Coping Trick: The essence of this trick is to identify those statements that you are confident are correct and those that you are confident are incorrect. This is best done by memory. There will be a few standard statements that you know to be correct and a few transpositions that you know to be incorrect. Memorise these. Then there is the huge range of statements in between. These may be correct or incorrect. The prosecutor may have transposed in his or her head and is trying to get you to say, what he thinks is a more simple statement. That is his fault not yours (if you are a forensic scientist reading this). He should have read and studied more. In this circumstance I suggest you say something like:

"I have been taught to be very careful with probabilistic statements. Subtle misstatements have led to appeals in the past. I am unsure whether your phrasing is correct or incorrect. However I can give some statements that I know are correct."

These will include the numerical statement of type 1, 2, or 3 given above or the verbal statements given in table 2.3.

Of course care by the scientist is no guarantee that the jury, judge or press will not make the transposition themselves. For instance Bruce Weir had gone to great trouble with the wording in the report for his testimony in the OJ Simpson case. Weir was careful and correct in his verbal testimony as well. As an example he reported that there was a 1 in 1,400 chance that the profile on the Bronco centre console would have this DNA profile IF it had come from two people other than Mr Simpson and Mr Goldman. This was transposed by Linda Deutsh of the Associated Press (June 26th, 1995) to "*a chance of 1 in 1,400 that any two people in the population could be responsible for such a stain*." To quote Professor Weir: "*It is incumbent on both the prosecution and defence to explain the meaning of a conditional probability of a DNA profile*."[835]

I found another transposition in an interesting place. Horgan[415] was warning about errors in the Simpson case and went on to commit the prosecutor's fallacy whilst explaining the error of the

defender's fallacy! "*Given odds of 1 in 100,000 that a blood sample came from someone other than Simpson, a lawyer could point out that Los Angeles contains 10 million people and therefore 100 other potential suspects. That argument is obviously specious…*" All the students in the 2003 (University of Auckland, NZ) Forensic Science class spotted the error when given it as an assignment!

Example from the 2003 Exam

d. The get away car hit a pole on leaving. You have compared the paint on the pole and from the car using microscopic and chemical analyses and the result is a match. You have testified that the probability of obtaining this paint result IF the paint came from a random car is 1 in 1000. The prosecutor asks you: "Do you mean there is only a 1 in 1000 chance that another car could have left this paint?" Please give your reply as you might in court. [4 marks]

A reply featuring the transposed conditional here. May mention Stella's spotting trick and Ian's coping trick.

Example from the 2007 exam

(d) The **Wikipedia** entry for a case in Western Australia is given below. The case was a paternity case, and the paternity index was 3134.

*"**Robert Bropho** (born 1930) is an indigenous Australian activist in Perth, Western Australia. He was leader of the Swan Valley Nyungah Community settlement for over 40 years. He organised the protest against redevelopment of the Swan Brewery, and was involved in the repatriation of Yagan's head. In 1986, he published Fringedweller.*

In 2003, the Swan Valley Nyungah Community settlement was closed amidst claims of widespread sexual abuse, rape and substance abuse, after a 15 year old girl named Susan Taylor committed suicide. Taylor's mother, Lena Spratt, accused Bropho of sexual misconduct against herself and her daughter. In September 2004, Bropho was found not guilty of two charges of raping a teenage girl nearly thirty years before, after a judge ruled inadmissible DNA evidence that alleged Bropho to be 3,134 times more likely to be the father of the woman's child than a random person. In December 2005, he was found guilty of indecently dealing with a girl under the age of thirteen, and sentenced to twelve months' jail. On January 30, 2006, he was be tried on a similar set of charges relating to another young girl who lived at the settlement, but apparently acquitted. His appeal against his conviction was rejected by the Court of Criminal Appeal in June 2006."

Is the highlighted statement most near to a statement of the paternity index, probability of exclusion or probability of paternity? Please explain your logic.

(5 marks)

Despite the fact that the witness gave the PI only the Wikipedia statement is a statement of a posterior odds on paternity. This is probably nearest to a probability of paternity. The Wikipedia writer has either assumed prior odds of 1:1 which is very odd in a criminal case or more likely transposed the conditional.

Two laws of heredity have been developed from Mendel's work. In modern times they are often phrased with the benefit of hindsight. We now know the chromosomal basis of inheritance associated with meiosis. However, at the time that Mendel wrote, none of this was known. An elegant phrasing of Mendel's laws without over-reliance on modern terminology is given by Thompson.[768] We follow her treatment here:

1. **The law of segregation**. Each individual has two 'factors' controlling a given characteristic, one being a copy of a corresponding factor in the father of the individual and one being a copy of the corresponding factor in the mother of the individual. Further, a copy of a randomly selected one of the two factors is copied to each child, independently for different children and independently of the factor contributed by the spouse.

2. **The law of independent assortment**. The factor copied from one pair is independent of the factor copied from another factor pair.

Modern molecular biology allows us to see the basis for these laws in the segregation of chromosomes and their recombination into a zygote. The human genome is diploid. It has a normal complement of 46 chromosomes arranged into 22 pairs of **autosomes** and a single pair of sex chromosomes (XY), the **gonosomes**. The somatic cells divide mitotically to maintain their diploid status whereas the sex cells (gametes) are produced by meiotic divisions and are haploid. During meiosis one of each of the pairs of the homologous chromosomes is randomly partitioned to the ovum or spermatozoon. In addition, there are recombination events that 'shuffle' the genetic material further still. At fertilisation the union of an ovum and a single spermatozoon restores the diploid chromosomal constitution and in doing so ensures that the embryo receives a random assortment of genes, half provided by one biological parent and the remaining half from the other biological parent (see figure 10.1). Mendel's laws form the basis of familial testing.

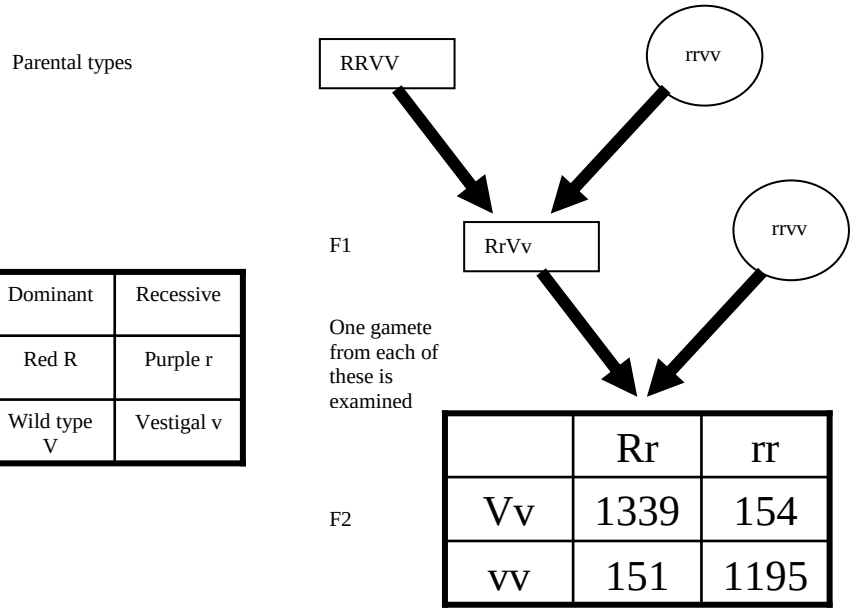
Punnett square

The law of segregation is often expressed by the use of a Punnett square (Punnett 1927). For the above example we would write the genotype of one parent across the top of the square, separating the alleles, and the other parent down the left hand side, also separating the alleles. The genotypes of the children are formed by taking the allele from the column and the row. Each combination is equiprobable if Mendel's first law applies and hence each combination occurs with probability ¼.

		Genes from one parent	
		a	b
Genes from the other parent	c	ac	bc
	d	ad	bd

An example of a Punnett square.

RECOMBINATION



	Dominant	Recessive
Eye colour	Red R	Purple r
Wing	Wild type V	Vestigal v

	Rr	rr
Vv	1339	154
vv	151	1195

Morgan took parental types RRVV and rrvv and crossed them. The F1 generation was therefore RrVv. He "backcrossed" these to the recessive rrvv. If the two loci for eye colour and wing assort independently (Meldel's 2nd law) then we expect the F1 generation to make four types of gametes in equal number. These four types are RV, Rv, rV, and rv. But in fact Morgan observed more of RV and rv. This was because the loci were close on the same chromosome.

$$\text{recombination fraction } (Rc) = \frac{\#\text{ recombinants}}{\#\text{ gametes examined}}$$

Definition: one map unit (m.u.) = **recombination fraction x 100.**

In honor of the work performed by Morgan, **one m.u. = one centimorgan (cM)**.

In the Morgan backcross there are 151+154 recombinants and 151+154+1339+1195 gametes examined. Hence $Rc = 0.1074$ this equates to 10.74cM

Map distance, recombination fraction, and Kosambi distance by CM Triggs. A genetic map distance of 1 Morgan is that distance such that one cross-over is expected to occur within it per gamete per generation. Typically data is expressed in centiMorgans (cM) and in humans 1cM is assumed to equal approximately 1000kb.

The simplest relationship between distance and recombination fraction is due to Haldane.[383] Consider two loci, *A* and *B*, and denote the genetic distance between them as *x*, and their recombination fraction as *Rc*.

$$Rc = \frac{1}{2} \times \left(1 - e^{-2x}\right) \quad \text{.....Haldane.}$$
$$e \approx 2.72$$

Kosambi took into account the fact that the strands of the DNA molecule are to some extent rigid and hence that the occurrence of a crossover will inhibit the possibility of a second nearby recombination event. He gives the relationship between the recombination fraction, *R*, and the map distance by:

$$Rc = \frac{1}{2} \times \frac{1-e^{-4x}}{1+e^{-4x}} \text{ .....Kosambi.}$$

Haldane, J.B.S 1919. The combination of linkage values, and the calculation of distance between linked factors. J. Genet. 8:299-309.

Kosambi, D.D. 1944. The estimation of map distance from recombination values. Ann. Eugen. 12:172-175.

EXAMPLE

Assume that the map distance was 30centimorgans the recombination fraction is

$$Rc = \frac{1}{2} \times \left(1-e^{-2\frac{30}{100}}\right) \text{ Haldane}$$
$$= 0.226$$

$$Rc = \frac{1}{2} \times \frac{1-e^{-4.\frac{30}{100}}}{1+e^{-4.\frac{30}{100}}} \text{ Kosambi}$$
$$= 0.269$$

Example (in the style of) the 2008 exam

(b) The loci X and Y are separated by 34.7 centiMorgans (cM) on chromosome 5. A certain man is genotype ab at X and cd at Y. His parents were aa and bb at X and cc and dd at Y. He has children with a woman who is bb at X and dd at Y.

(i) Please give the expected recombination fraction, *R*c. (2 marks)

$$Rc = \frac{1}{2} \times \left(1-e^{-2\times\frac{34.7}{100}}\right) = 0.25$$

(ii) Fill in the following table indicating what fraction of his children are expected to be each genotype. (2 marks)

		X	
		ab	bb
Y	cd		
	dd		

You may use Haldane's equation $Rc = \frac{1}{2} \times \left(1-e^{-2x}\right)$ where x is the distance in Morgans and *R*c is the recombination fraction.

		X	
		ab	bb
Y	cd	0.375	0.125
	dd	0.125	0.375

GENETIC DRIFT

Genetic drift is the accumulation of purely random changes in relative abundance of allele frequencies in a population

POPULATION BOTTLENECK

A population bottleneck (or genetic bottleneck) is an evolutionary event in which a significant percentage of a population or species is killed or otherwise prevented from reproducing, and the population is reduced by 50% or more, often by several orders of magnitude.

FOUNDER EFFECT

The establishing a new population by a small number of individuals

EXAMPLE.

Polydactyly (extra fingers and toes, a symptom of Ellis-van Creveld syndrome) is more common in Amish communities than in the US population at large. It is caused by a recessive allele. The Amish community was founded by about 200 individuals. The Old Order Amish Ellis-van Creveld syndrome has been traced back to one couple, Samuel King and his wife, who came to the area in 1744. Today it is many times more common in the Amish population (0.066 of live births) than in the American population at large (1 case per 60,000 live births).

Is this founder effect, bottleneck, or drift?

Allele frequency at foundation: Two individuals both carrying one recessive allele out of 200 individuals. Frequency $p = \frac{2}{400} = 0.005$

Allele frequency today: For a recessive allele you need two copies to have the syndrome. So $p^2 = 0.066 \rightarrow p = 0.25$

Frequency in the US: $p^2 = \frac{1}{60,000} \rightarrow p = 0.004$

So the founding population was not too far from a "normal" population (the US today). It is therefore drift since 1744. BTW founder effect is really a type of genetic bottleneck

GENE FLOW

Gene flow is the exchange of genes between populations, which are usually of the same species. It may occur with or without the physical movement of individuals

DRIFT-MIGRATION EQUILIBRIUM

Since drift and migration work in opposite directions they can form an equilibrium

$$\theta \approx \frac{1}{1+4Nem} \qquad Ne = \frac{4NmNf}{Nm+Nf}$$

Ne = the effective size of the population
Nm, Nf are the numbers of adult males and females

m is the fraction of the population replaced by migrants

Example: Estimate θ assuming migration drift equilibrium from these data:

- Australian intertribal marriage was of the order of 14% of marriages pre-contact
- Assume sex ratio 3:2 M:F
- Assume tribal size 500 of which 250 are juveniles or old
- Estimate *N*e and hence θ please

There are 150 adult males and 100 adult females hence $Ne = \frac{4 \times 150 \times 100}{150 + 100} = 240$

If 14% or marriages are intertribal we assume this is 7% of people out and 7% in. Hence $m = 0.07$

$$\theta = \frac{1}{1 + 4 \times 240 \times 0.07} = 0.015$$

Example from the 2008 exam

(v) Two of this population found a new population. They had genotypes AB and AB. 100 years later the population has grown to 100 from the original 2. The new population has genotype frequencies

Genotype	Count number of individuals
AA	4
AB	32
BB	64

Please describe what has happened to this population using the correct population genetic description (2 marks)

Example from the 2009 exam

Q5a

At T=0. A population on an island has a population of 9000. A disease occurs in 1 in 900 people in this population. The disease is caused by a recessive gene.

At T=1 a natural disaster drops the population to 10 one of who is a heterozygotic carrier of the disease gene. 20 generations later (T = 21) 1 in 20 people are showing the disease.

i)What is the frequency of the disease gene at T=0, T=1, and T = 21? 3 marks
ii)What evolutionary phenomena have occurred? 2 marks

Model answers:
T=0 0.0333
T=1 0.05
T=21 0.224
Bottleneck and drift

5b An island has a population of 1200. On average 50% of the individuals are juvenile or old. The male to female ratio is 2:1. On average 8% of marriages are with neighbouring islands. What value do we expect for θ if a drift migration equilibrium has formed. You may use:

$$\theta \approx \frac{1}{1 + 4Nem} \qquad Ne = \frac{4NmNf}{Nm + Nf}$$

Ne = the effective size of the population
Nm, Nf are the numbers of adult males and females
m is the fraction of the population replaced by migrants

Ne = 533

$\theta = 0.0116$ 3marks

What will happen to θ if the population grows to 12,000 and the drift migration equilibrium re-establishes.

$\theta = 0.00117$ 2marks

POPULATION GENETIC MODELS

We will discuss three models in common use. The product rule, recommendation 4.1 of NRC II and recommendation 4.2 of NRC II.

PRODUCT RULE

This is the simplest of the available population genetic models. It is based on the Hardy-Weinberg law and the concept of linkage equilibrium.[805,806]

HARDY-WEINBERG LAW

This concept was first published in 1908[392,826] although simplified versions had been published previously.[151,611,878] This thinking developed naturally following the rediscovery of Mendel's work.[546] It concerns the relationship between allele probabilities and genotype probabilities at one locus. In essence the Hardy-Weinberg law is a statement of independence between alleles at one locus.

The Hardy-Weinberg law states that the single locus genotype frequency may be assigned as the product of allele probabilities

$$P_i = \begin{cases} p_{i1,}^2 & A_{i1} = A_{i2} \\ 2p_{i1}p_{i2,} & A_{i1} \neq A_{i2} \end{cases} \quad \text{.....equation 3.1 for alleles } A_{i1}, A_{i2} \text{ at locus } i.$$

This will be familiar to most in the form

$$\begin{cases} p^2 & \text{homozygotes} \\ 2pq & \text{heterozygotes} \end{cases}$$

The assumptions that make the Hardy-Weinberg law true are that the population is infinite, randomly mating and that there are no disturbing forces. Inherent in this law is the assumption of independence between genotypes.

The assumption of random mating assumes that the method of selection of mates does not induce dependence between genotypes. What is suggested is that geography, religion or some other socio-economic factors induce dependence. .

There are, however, a number of factors that can change allele proportions. These are referred to as disturbing forces. The term is derived from the fact that they change genotype proportions from those postulated by HWE. These factors include selection, migration, and mutation.

B. LINKAGE AND LINKAGE EQUILIBRIUM

Hardy-Weinberg equilibrium describes a state of independence between alleles at one locus. Linkage equilibrium describes a state of independence between alleles at different loci.

The same set of assumptions that gives rise to Hardy-Weinberg equilibrium plus an additional requirement that an infinite number of generations has elapsed also lead to linkage equilibrium. This result was generalised to three loci by Geiringer,[332] and more generally to any number of loci by Bennett.[54]

It is worthwhile discussing the difference between linkage equilibrium and linkage, as there is an element of confusion about this subject amongst forensic scientists. Linkage is a genetic phenomenon and describes the situation where one of Mendel's laws breaks down. It was discovered in 1911 by Morgan[555,556] working on *Drosophila.* The discovery was a by product of his team's studies of inheritance that had largely led to the confirmation of the chromosomal theory of inheritance. The first paper on gene mapping appeared in 1913.[740]

Specifically the phenomenon of linkage describes when alleles are not passed independently to the next generation. The physical reason for this phenomenon had been identified by 1911 and

related to the non-independent segregation of alleles that are sufficiently close on the same chromosome.[597]

The state of linkage can be described by the recombination fraction or by the distance between two loci. Typical data for distance may be expressed in centiMorgans (cM) or in physical distance in bases. In humans 1cM is assumed to equal approximately 1000kb.

The physical distance may be converted to a recombination fraction by standard formulae.[4] Recombination fractions tend to be different for each sex. Distances may be given separately or sex-averaged.

Linkage disequilibrium is a state describing the relationship between alleles at different loci. It is worthwhile pointing out that linkage disequilibrium can be caused by linkage or by other population genetic effects such as population subdivision

If the population is in linkage equilibrium then a multilocus genotype probability (P) may be assigned by the product of single locus genotype probabilities (P_i).

$$P = \prod_i P_i \quad \text{.....equation 3.2}$$

The Wahlund effect

This leads us to the classical consideration of the Wahlund principle.[801] Assume that a certain area is made up of two or more subgroups that breed within each group but not to any large extent between the two groups. Further assume that there are some allele probability differences between these groups. Then even if the subpopulations themselves are in Hardy-Weinberg equilibrium the full population will not be. An example is given in table 3.2.

First we note that the mixed population is not in Hardy-Weinberg equilibrium even though each subpopulation is. Next we note the classical Wahlund effect that all the probabilities for homozygotes are increased above Hardy-Weinberg expectation. The total heterozygote probabilities are generally decreased although individual heterozygotes may be above or below expectation. Note that in this example two of the heterozygotes are below expectation whereas one is above. The total for all the heterozygotes will always be down (which is really the same as saying the total of the homozygotes is always up).[267,836]

Table 3.2: An example of the Wahlund effect

Allele	*a*	*b*	*c*
Subpopulation 1	0.7	0.2	0.1
Subpopulation 2	0.2	0.1	0.7

Genotype	Subpopulation 1	Subpopulation 2	1:1 Mix	Hardy-Weinberg expectation
aa	0.49	0.04	0.2650	0.2025
bb	0.04	0.01	0.0250	0.0225
cc	0.01	0.49	0.2500	0.1600
ab	0.28	0.04	0.1600	0.1350
ac	0.14	0.28	0.2100	0.3600
bc	0.04	0.14	0.0900	0.1200

[4]See Chapter 1 footnote iii

Example from the 2003 exam: A survey of genotype counts at a certain locus with three alleles was undertaken. Below are the genotype counts for sample of 1000 from the English and Irish populations.

Genotype	English	Irish
15,15	90	10
15,16	300	120
15,17	120	60
16,16	250	360
16,17	200	360
17,17	40	90
	1,000	1,000

A certain area, "Tasmania" is populated by 4000 English and 1000 Irish people. Please fill in the following table for "Tasmania"

Genotype	English	Irish		Tasmania actual	Tasmania expected
15,15	90	10		370	338
15,16	300	120		1,320	1,352
15,17	120	60		540	572
16,16	250	360		1,360	1,352
16,17	200	360		1,160	1,144
17,17	40	90		250	242
	1,000	1,000		5,000	5,000
	Allele probabilities→		15	0.26	
			16	0.52	
			17	0.22	

i. In the "actual" column please place the counts formed by the total population of 5000 comprising 4000 English and 1000 Irish people.

ii. Please give the allele probabilities for "Tasmania actual"

iii In the expected column please place the expected counts if the substructure were ignored and Hardy-Weinberg equilibrium was assumed. [6 marks]

2 marks for each part i-iii. If an error was made early but then the resulting results were correct I recalculated and gave part marks for the correct portions.

3c Please use this example to describe the Wahlund effect. [3 marks]

Mention of all homs above expectation. Total hets below but some may be above. Subtract ¼ mark if there is no explicit mention that some hets may be up and some down. Must mention that all homs are up and total hets down or nil marks.

Example from the 2006 Exam
4. Please answer part a and part b.

(a) In a certain area of Belgium a sample of 5000 people was taken. The following are the sample results:

	Belgium
aa	370
ab	280
ac	1020
ad	560
bb	80
bc	440
bd	320
cc	730
cd	880
dd	320

Please calculate the allele probabilities for the a, b, c, and d alleles for this area of Belgium. (2 marks)

Pr(a) = 0.26 Pr(b) = 0.12 Pr(c) = 0.38 Pr(d) = 0.24

What are the expected genotype probabilities if this area is in Hardy-Weinberg equilibrium? (5 marks)

	Belgium	Expected under HW
aa	370	338
ab	280	312
ac	1020	988
ad	560	624
bb	80	72
bc	440	456
bd	320	288
cc	730	722
cd	880	912
dd	320	288
		0
		5000

(iii) What are the assumptions that lead to Hardy-Weinberg equilibrium? (2 marks)

Infinite population, random mating, no selection, migration or mutation

(iv) Is the population of this area of Belgium in Hardy-Weinberg equilibrium? (1 mark)

No

NRC II RECOMMENDATION 4.1.

NRC II recommendation 4.1 offered a correction for Hardy-Weinberg disequilibrium caused by the Wahlund effect. It was suggested that a correction upwards in frequency be applied to correct for the expected upward bias produced by population subdivision. Further that this correction should be applied only to homozygotes. No correction was recommended for heterozygotes since, on average these should have a downward bias (recall that individual heterozygotes may be displaced from expectation in either direction). This comment is generally true for the event of population subdivision but would be untrue for populations undergoing admixture. In admixing populations the number of heterozygotes is likely to be elevated.

The recommendation suggests:

$$P_i = \begin{cases} p_{i1}^2 + p_{i1}(1-p_{i1})F & A_{i1} = A_{i2} \\ 2p_{i1}p_{i2,} & A_{i1} \neq A_{i2} \end{cases} \quad \text{……..equation 3.3}$$

where F is the within person inbreeding coefficient not the between person inbreeding coefficient, θ, as written in NRC II.

This recommendation is a logical way of correcting for Hardy-Weinberg disequilibrium but makes no attempt to correct for linkage disequilibrium. It will suffer from the same approximations that are revealed in Table 3.2 for the 1:1 mix from genotypes. Hence it will still have a very mild tendency to underestimate multilocus genotype probabilities.

Curran et al. tested recommendation 4.1 by comparing this assignment with the "Gold Standard Profile Frequency" for a population with a true inbreeding coefficient θ =0.03 created by simulation. This is reproduced in figure 3.4. In this simulation 54.4% of values are less than 1 (reduced from 64.7% for no correction). We see that this estimator still has a small prosecution bias and some undesirable variance properties.

THE SUBPOPULATION FORMULAE

If it is difficult to calculate the genotype probability in the population due to the effects of population subdivision, can we calculate it in the subpopulation of the suspect? We note that the subpopulation of the suspect may not be known, may not be easily defined, and almost certainly has not been sampled.

A potential solution has been offered by Balding and Nichols and has found widespread acceptance both in the forensic and the legal communities. These formulae[29,36,41,267,585] calculate the conditional probability of a second profile matching the stain from the subpopulation of the suspect given the profile of the suspect.

These formulae follow from a formal logic given initially by Balding and Nichols and appearing as equations 4.10 in NRC II and 4.20 in Evett and Weir but they date back to the work of Sewall Wright[873] in the 1940's. A reasonably gentle derivation appears in Balding and Nichols.[39]

$$P_i = \begin{cases} \dfrac{\left[3\theta + (1-\theta)\,p_{i1}\right]\left[2\theta + (1-\theta)\,p_{i1}\right]}{(1+\theta)(1+2\theta)}, & A_{i1} = A_{i2} \\ \dfrac{2\left[\theta + (1-\theta)\,p_{i1}\right]\left[\theta + (1-\theta)\,p_{i2}\right]}{(1+\theta)(1+2\theta)}, & A_{i1} \neq A_{i2} \end{cases}$$

$$P = \prod_i P_i \quad \text{……..equation 3.4}$$

Example from the 2006 Exam

(b) This area of Belgium is thought to be populated by two subpopulations: the Walloons and the Flems. They have an inbreeding coefficient $\theta = 0.03$. In a certain case a suspect is identified. He has genotypes ab. The stain at the scene is genotype ab.

(i) Please calculate the probability of the genotype ab using the product rule. (2 marks)

$$2\Pr(a)\Pr(b) = 2 \times 0.26 \times 0.12 = 0.0624$$

(ii) Please calculate the probability of the genotype ab using NRC II recommendation 4.1. (2 marks)

The same (no correction for heterozygotes.

(iii) Please give the formula for the probability of the genotype ab using NRC II recommendation 4.2 and evaluate it. (3 marks)

$$\frac{2\left[\theta + (1-\theta)\Pr(a)\right]\left[\theta + (1-\theta)\Pr(b)\right]}{(1+\theta)(1+2\theta)} = \frac{2\left[0.03 + (0.97 \times 0.26)\right] \times \left[0.03 + (0.97 \times 0.12)\right]}{1.03 \times 1.06} = 0.0757$$

(iv) What is the expected performance with respect to conservativeness of the product rule, NRC 4.1 and 4.2 in this instance? (3 marks)

Not taught in 2009

Shortcut rules

The shortcut rules are demonstrated by way of examples given below. These rules are not really 'derivations' but are a set of rules that allow the answer to be written down. With practice this becomes second nature. We begin by writing the probability in the conditional form. In front of the conditioning bar we place the genotype(s) of the 'possible offender(s)'. Behind the bar we place the conditioning genotype(s). This should always include the suspect but in some circumstances other profiles may also be included here. This has become an area of some debate which is covered in a short section later in the chapter.

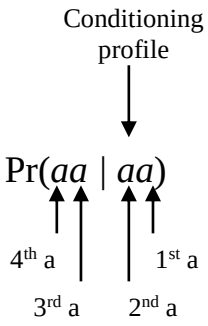


Figure 7.4 A diagrammatic representation to assist evaluation using the shortcut rules.

Example 7.8 The calculation of Pr(*aa*|*aa*).

Although our purpose is to demonstrate the application of this process to mixed stains it is easiest to start with a simple example of a case where the stain at the scene is unmixed and shows the genotype *aa*. The suspect is *aa*. Hence we see that the only genotype for 'possible offenders' is *aa* and the only potential conditioning profile is the suspect, also *aa*. Accordingly in this example we consider the calculation of the conditional probability Pr(*aa*|*aa*) shown figuratively in Figure 7.4. The following three steps are required to obtain the formula.

Apply a factor of 2 if the 'possible offender' is heterozygous. The 'possible offender' will be the term in front of the conditioning bar. In this example the 'possible offender' is the homozygote *aa* therefore no factor of 2 is required.

Counting from the back towards the front; label each allele as the first of this type seen, second of this type seen and so on. Replace each of the possible offender's alleles with the terms given in Table 7.10. It is necessary to proceed from one or other end of the offender's genotype. For instance in the calculation of Pr(*aa*|*aa*) we see that the homozygote *aa* in front of the conditioning bar is treated as the 3rd and 4th *a* alleles.

Table 7.10 The conversion of terms using the shortcut rules.

1st allele *a*	$(1-\theta)p_a$
2nd allele *a*	$\theta+(1-\theta)p_a$
3rd allele *a*	$2\theta+(1-\theta)p_a$
4th allele *a*	$3\theta+(1-\theta)p_a$
…..	

Divide by a correction term based on the number of alleles in front of and behind the conditioning bar shown in Table 7.11

Table 7.11 The correction terms

2 alleles in front, 2 behind	$(1+\theta)(1+2\theta)$
2 in front, 4 behind	$(1+3\theta)(1+4\theta)$
2 in front, 6 behind	$(1+5\theta)(1+6\theta)$
4 in front, 2 behind	$(1+\theta)(1+2\theta)(1+3\theta)(1+4\theta)$
4 in front, 4 behind	$(1+3\theta)(1+4\theta)(1+5\theta)(1+6\theta)$
4 in front, 6 behind	$(1+5\theta)(1+6\theta)(1+7\theta)(1+8\theta)$
N in front, *M* behind	$[1+(M-1)\theta]\ldots$ $[1+(N+M-3)\theta][1+(N+M-2)\theta]$

This yields the familiar formula $\Pr(aa \mid aa) = \dfrac{(3\theta+(1-\theta)p_a)(2\theta+(1-\theta)p_a)}{(1+\theta)(1+2\theta)}$

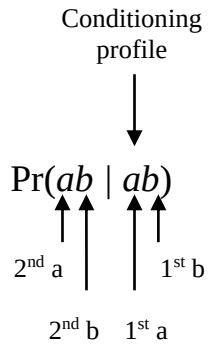


Figure 7.5 A diagrammatic representation to assist evaluation using the shortcut rules.

Example 7.9 The calculation of Pr(*ab*|*ab*).

Consider the calculation of Pr(*ab*|*ab*) shown diagrammatically in Figure 7.5. Application of the rules leads quickly to the familiar formula

$$\Pr(ab \mid ab) = \frac{2(\theta + (1-\theta)p_a)(\theta + (1-\theta)p_b)}{(1+\theta)(1+2\theta)}$$

Example 7.10

As a more practical example consider the following where the complainant (of race 1) has been genotyped as *ab*, the suspect (of race 2) has been genotyped as *cc*, and a semen-stained swab taken from the complainant after an alleged assault has been genotyped as *abc*. In the absence of any quantitative information the genotype of the offender could be *ac*, *bc* or *cc*.

Complainant	Race 1	Typed as *ab*
Suspect	Race 2	Typed as *cc*
Swab		Typed as *abc*

It is unreasonable to assume that the complainant and the suspect are from the same subpopulation if they are of different races. This assumption follows from a rigid application of a hierarchical population/sub population approach. However subpopulations from different races could share alleles that are identical by descent (IBD) by recent admixture, in which case this simplification may not be valid. Following the arguments of Nichols and Balding,[47] the suspect and offender are assumed to be from the same subpopulation.

The likelihood ratio uses the probabilities of the offender's type conditional on the suspect's type (the complainant's type is ignored as having come from a different population):

$$LR = \frac{1}{\Pr(ac \mid cc) + \Pr(bc \mid cc) + \Pr(cc \mid cc)}$$

since $\Pr(ac \mid cc) = \dfrac{2(1-\theta)p_a\left[2\theta + (1-\theta)p_c\right]}{(1+\theta)(1+2\theta)}$

$$\Pr(bc \mid cc) = \frac{2(1-\theta)p_b\left[2\theta + (1-\theta)p_c\right]}{(1+\theta)(1+2\theta)}$$

$$\Pr(cc \mid cc) = \frac{\left[3\theta + (1-\theta)p_c\right]\left[2\theta + (1-\theta)p_c\right]}{(1+\theta)(1+2\theta)}$$

$$LR = \frac{(1+\theta)(1+2\theta)}{(2\theta + (1-\theta)p_c)(3\theta + (1-\theta)(2p_a + 2p_b + p_c))}$$

Substitution of $\theta = 0$ recovers the product rule formulae given in Table 7.1 $LR = \dfrac{1}{p_c(2p_a + 2p_b + p_c)}$ and provides a useful check.

2. When should a genotype be used in the conditioning?[5]

The subpopulation model works best when those people who share the same subpopulation as the suspect are used in the conditioning. There are many complicating factors in this. These include

[5] This matter was brought to our attention by a senior caseworker in New Zealand, Sue Vintiner. It has been constructively discussed in meetings in New Zealand and in conversations with Robert Goetz, Manager of the Forensic Biology Laboratory of the Division of Analytical Laboratories, NSW, Australia.

- The subpopulation of the suspect may both undefinable and unknown.
- The subpopulation of any other typed person may be both undefinable and unknown.

Clearly the suspect is a member of his or her own subpopulation whether or not we know that or can define it. But who else is? In many cases this is unanswerable. The inclusion of additional genotypes in the conditioning if they are not members of the suspect's subpopulation essentially adds an unwelcome random element. Such an addition is not expected to improve the estimation process at all but rather adds variance about the true value. The addition of such people tends to give a more conservative LR when the person and the suspect share many alleles. It tends to give a less conservative LR when the person and the suspect share few or no alleles. It had been supposed that the addition of random persons was conservative **on average**. We are uncertain whether this is true but even if true it applies on average over a number of cases rather than in each case. Accordingly we consider that the addition of "random" genotypes to the conditioning may make the LR more or less conservative but does not improve the process of obtaining the best estimate.

The effect of adding random genotypes is to randomise the answer.

As a first approximation, we suggest that only those persons **known** or **reasonably assumed** to share the subpopulation of the suspect should be added to the conditioning. This knowledge will very rarely be available in casework and hence most often only the suspect's genotype will appear behind the conditioning.

If the forensic scientist wishes to report the more conservative estimate we cannot think of anything better at this time than calculating the likelihood ratio both ways and reporting the smaller.

Example: A crime occurs in a small rural village in Switzerland. The crime stain is genotype ab. The suspects are all local men from families that have lived in the area for a long time. Suspect 1 is genotype ab, suspect 2 ac and suspect 3 bd.

Who is behind the bar? I think all three can be assumed to be from the same subpopulation.

So we want:

$$\Pr(ab \mid abacbd) = \frac{2(2\theta + (1-\theta)P_a)(2\theta + (1-\theta)P_b)}{(1+5\theta)(1+6\theta)}$$

EXAMPLE FROM THE 2007 EXAM

3. Please answer both parts of part (a), all parts of part (b), and part (c).

(a) (i) What are the assumption that lead to Hardy-Weinberg equilibrium?
(2 marks)

Infinite population, random mating, no migration, mutation or selection

(ii). What happens when you mix two different populations?
(2 marks)

The Whalund effect. More homozgyotes than expected fewer total heterozygotes although each heterozygote genotype may be up or down.

(b) (i) In a case we seek the probability of an ab heterozygote from a subpopulation where we have observed the suspect, ab, and one other individual who was ac. Please give the formulation for Pr(ab|aabc) using the method of recommendation 4.2.

First a	$(1-\theta)\Pr(a)$
Second a	$\theta+(1-\theta)\Pr(a)$
Third a	$2\theta+(1-\theta)\Pr(a)$
Fourth a	$3\theta+(1-\theta)\Pr(a)$

Two allele in front and two behind	$(1+\theta)(1+2\theta)$
Two allele in front and four behind	$(1+3\theta)(1+4\theta)$
Two allele in front and six behind	$(1+5\theta)(1+6\theta)$

$$\Pr(ab\,|\,abac)=\frac{2\left[2\theta+(1-\theta)\Pr(a)\right]\left[\theta+(1-\theta)\Pr(b)\right]}{(1+3\theta)(1+4\theta)}$$

(4 marks)

(ii) Define θ when used in this context as if you were explaining it in court. (2 marks)

θ can be viewed as a measure of relatedness between two different individuals or as the genetic distance between the subpopulations

(iii) Please set $\theta=0$ in the formula. What do you obtain? Why? (2 marks)

$\Pr(ab\,|\,abac)=2\Pr(a)\Pr(b)$ which is the product rule. This occurs because $\theta=0$ means that there is no distance between subpopulations and hence only one population.

EXAMPLE FROM THE 2008 EXAM

(b) Two exchange students Mr A and Mr B from a remote isolated village, Gioja, in Europe are studying at Auckland University. They end up on the town with two Kiwi friends Mr C and Mr D who they have met at the gym.

At a pub near the university later that night a man is seen to drunkenly strike the barman. In the scuffle he bleeds and runs away. The blood at the scene is typed at the vWA locus and is type 14,18

Two days later the Police investigate Messrs A, B, C and D. They all willingly give DNA samples. The types are

Mr A	14,18
Mr B	14,19
Mr C	15,18
Mr D	16,16

(i) Please produce the estimate for the probability of a 14,18 genotype using the product rule, NRC recommendation 4.1, and NRC recommendation 4.2. You may use $f_{14} = 0.10$, $f_{15} = 0.12$, $f_{16} = 0.15$, $f_{18} = 0.08$, $f_{19} = 0.06$, $F = 0.03$, $\theta = 0.03$

(5 marks)

Product rule $2 \times 0.10 \times 0.08 = 0.016$

4.1 same (no correction for heterozygotes)

4.2 $\Pr(14,18 \mid 14,18,14,19) = \dfrac{2\left[2\theta + (1-\theta) f_{14}\right]\left[\theta + (1-\theta) f_{18}\right]}{(1+3\theta)(1+4\theta)} = 0.0224$

(ii) What are the assumptions of the product rule? (2 marks)

Hardy-Weinberg and Linkage equilibrium. Good to spell out the assumptions of HW and LE as well.

(iii) Which one would you use in a criminal trial in New Zealand, and why?

(3 mark

Not taught in 2009

Example from the 2009 exam

Q2b A crime occurs are a blood stain is left by the offender at a scene. The genotype of the blood stain is type aa. A group of four men become the suspects. All four are members of a small community from Europe. The genotypes of the four men are

Suspect 1: aa
Suspect 2 ac
Suspect 3 bd
Suspect 4 cd

You may use Pr(a) = 0.02, Pr(b) = 0.10, Pr(c)= 0.10, Pr(d)=0.10, $F = 0.03$, $\theta = 0.03$

Using the data above please give the estimate of the frequency of this genotype using the product rule and using NRC II recommendation 4.1? 3marks

Product rule: 0.0004
NRC 4.10.000998

When we use NRC II recommendation 4.2 we need to consider the conditional probability of the genotype of the offender given the alleles we know came from the subpopulation of the suspect. This is often set out in the form

$\Pr(\_\_ \mid \_\_\_\_\_)$. Please insert the correct alleles into this term and evaluate.

$$\Pr(aa \mid aaacbdcd) = \frac{(3\theta + (1-\theta)\Pr(a))(4\theta + (1-\theta)\Pr(a))}{(1+7\theta)(1+8\theta)} = 0.0102$$ 5 marks

For the hypotheses

Hp: Suspect 1 is the offender
Hd: A random man is the offender

What is the likelihood ratio using NRC II recommendation 4.2?

Ans: 98 2 marks

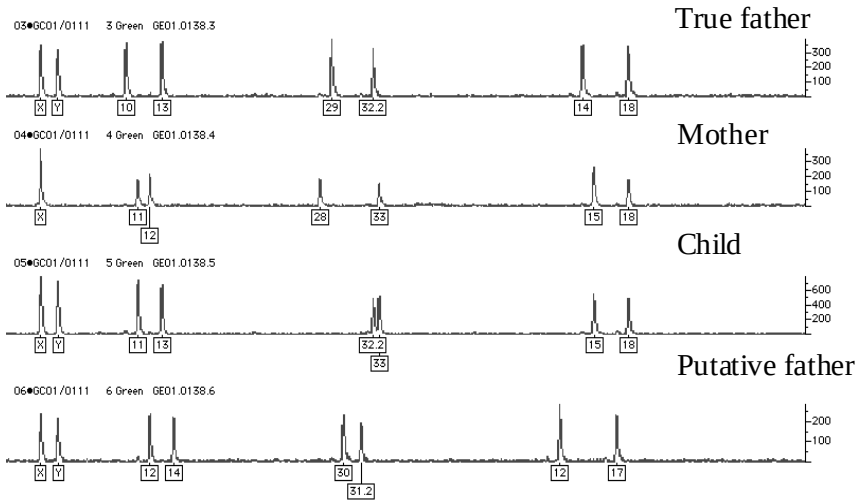


Figure 10.1 Profiles of mother, child, the true father and a putative father at four autosomal STR loci.

EVALUATION OF EVIDENCE

Three methods have been offered for the evaluation of parentage testing results. These are often termed the paternity index (*PI*), the probability of paternity, and an exclusion probability.[659,808] Strong support is given for the *PI* approach by many authorities including Evett and Weir[279] and the Paternity testing Commission of the International Society of Forensic Genetics.[567]

EXCLUSION PROBABILITY

Consider the most common case of parentage testing where we have a mother (M), child (C), and a man alleged to be the father (AF). These three persons have been typed and found to have the genotypes G_M, G_C, G_{AF}, respectively. The genotypes of the mother and the child define one (or in some cases one of two) paternal alleles at each locus.

An exclusion probability may be defined as "that fraction of men who do not possess the paternal allele or alleles." As such it is strongly akin to the exclusion probability in mixtures evaluation.

If the possible paternal alleles at a locus are $A_1 \ldots A_n$ (often there is only one possible paternal allele) then the exclusion probability at locus this locus (PE_l) is $PE_l = (1 - \sum_{i=1}^{n} \Pr(A_i))^2$ assuming Hardy-Weinberg equilibrium. The *PE* across multiple loci (*PE*) is calculated as $PE = 1 - \prod_l (1 - PE_l)$. For an extension to the consideration of relatives see Fung et al.[336]

We have previously discussed Dr Charles Brenner's[92] explanation of the shortcomings of the probability of exclusion. We follow his treatment again here.

Let us describe the evidence as:

1. The blood type of the mother,
2. The blood type of the child, and
3. The blood type of the alleged father.

From this information we can infer that:

4. The alleged father is not excluded.

Brenner points out that although statement 4 can be deduced from statements 1, 2 and 3, statement 3 cannot be deduced from 1, 2 and 4. Hence the use of statement 4 represents a loss of information. The exclusion probability is a summary of the evidence in 1, 2 and 4.

B. PATERNITY INDEX

The paternity index (*PI*) is a specialist term used in paternity testing to describe the likelihood ratio. Its structure is exactly as described for the likelihood ratio in Chapter 2 but has been used in paternity testing for longer than in other areas of forensic biology.[267] Hallenberg and Morling[395] reported that 73% of respondents in the year 2000 and 78% in 2001 used the paternity index or the probability of paternity to interpret parentage evidence. Consider the two hypotheses:

H_p: The alleged father is the true father.

H_d: The alleged father is the not the true father.

Hypothesis H_p represents one side of the allegation. In many paternity cases the action will be civil and it may not be appropriate to view this as the 'prosecution' hypothesis. Fortunately the same letter can stand for 'paternity'. Hypothesis H_d represents the other side of the allegation; similarly it may not be appropriate to view this as the 'defence' hypothesis.

If we consider some evidence, *E*, typically the genotypes of a child, the alleged father, and possibly the mother then Bayes' theorem informs us that:

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

The likelihood ratio term $\frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)}$ is usually written as *PI* and is the central term calculated under this approach.

Use of the product rule in the evaluation of the Paternity Index.

We have discussed the small bias inherent in the use of the product rule when population substructure exists. The method of Balding and Nichols[47] can be used to evaluate likelihood ratios, or Paternity Indices, for paternity duos and trios when population substructure exists.

When the Balding and Nichols' correction is applied to a whole race or when conservatively large values of θ are used this is thought to be an overcorrection which may err too much in one direction. This 'conservative' behaviour is considered desirable by some courts and scientists in criminal cases. However, this property of the subpopulation correction does not have such an obvious justification in civil cases.

PROBABILITY OF PATERNITY

Recall Bayes' theorem that states $\frac{\Pr(H_p \mid E)}{\Pr(H_d \mid E)} = PI \times \frac{\Pr(H_p)}{\Pr(H_d)}$. We see that the paternity index relates the odds on paternity prior to considering the genetic evidence to those after considering that evidence. As with any Bayesian treatment the posterior probability of paternity can be calculated from the paternity index and the prior odds. The prior odds relate to the probability of paternity based on the non-genetic evidence. This could include statements of the mother as to with whom she had intercourse, or evidence that may suggest that the alleged father was out of the country or in prison at the time of conception. Such evidence, if relevant and admissible, affects the prior odds.

However, it has become customary to set the prior odd to 1:1, that is to assign prior probabilities of 50% to both H_p and H_d, when calculating the probability of paternity. This assumption is hard to justify at the fundamental level{Robertson, 1995 #22;Good, 2001 #3144@ at pg 68 & 89-91} and must be seen simply as a pragmatic tool. It may be completely appropriate in many cases but equally may be totally inappropriate in others. It would seem wise, however, to make this assumption of equal prior odds explicit.

Utilising this assumption we see that $\frac{\Pr(H_p \mid E)}{\Pr(H_d \mid E)} = PI$ and hence that $\frac{\Pr(H_p \mid E)}{1 - \Pr(H_p \mid E)} = PI$ yielding $\Pr(H_p \mid E) = \frac{PI}{1 + PI}$.

We (and others) cannot support the assumption of prior odds despite its extensive use and rather advocate use of the *PI* alone.[67,659] This stance is taken by the Paternity Testing Commission of the International Society of Forensic Genetics:

> "*If the weight of the evidence is calculated, it shall be based on likelihood ratio principles. The paternity index, PI, is a likelihood ratio*"[568]

PATERNITY TRIOS: MOTHER, CHILD AND ALLEGED FATHER

We begin by considering at least two hypotheses. In the most common case these could be:

H_p: The alleged father is the true father, (and the mother is the true mother).

H_d: A random person who is not related to the alleged father is the true father (and the mother is the true mother).

The assumption that the person labelled as the mother is the true mother of the child is usually unstated. Although these two hypotheses are the most commonly used we note that they are not exhaustive as the random person may be a relative of the alleged father. This again suggests an alternative approach based on the general form of Bayes' theorem. Such an approach is not in use in any laboratory of which we are aware.

Typically then we require $PI = \frac{\Pr(G_C, G_M, G_{AF} \mid H_p)}{\Pr(G_C, G_M, G_{AF} \mid H_d)}$.

It is customary to decompose these probabilities using the third law of probability. Usually to evaluate the probabilities of the observing genotypes of individuals they are conditioned on the genotypes of their ancestors. For example:

$$PI = \frac{\Pr(G_C, G_M, G_{AF} \mid H_p)}{\Pr(G_C, G_M, G_{AF} \mid H_d)} = \frac{\Pr(G_C \mid G_M, G_{AF}, H_p)\Pr(G_M, G_{AF} \mid H_p)}{\Pr(G_C \mid G_M, G_{AF}, H_d)\Pr(G_M, G_{AF} \mid H_d)},$$

where the genotype of the youngest person, the child, is conditioned on the parents, as opposed to:

$$PI = \frac{\Pr(G_C, G_M, G_{AF} \mid H_p)}{\Pr(G_C, G_M, G_{AF} \mid H_d)} = \frac{\Pr(G_{AF} \mid G_M, G_C, H_p)\Pr(G_M, G_C \mid H_p)}{\Pr(G_{AF} \mid G_M, G_C, H_d)\Pr(G_M, G_C \mid H_d)}.$$

Both decompositions are, of course, formally equivalent mathematically. However the former is easier to evaluate. Thus we will work with the former decomposition.

It is customary to assume that the joint probability of observing the genotypes of the putative parents does not depend on the particular hypothesis, i.e.

$$\Pr(G_M, G_{AF} \mid H_p) = \Pr(G_M, G_{AF} \mid H_d) = \Pr(G_M, G_{AF}).$$

This assumption essentially states that the joint probability of observing the genotypes of the mother and alleged father are not conditioned on whether the alleged father is the true father or not. This is only true in the absence of any conditioning on the genotypes of any other children or descendants. Given this assumption the paternity index becomes

$$PI = \frac{\Pr(G_C \mid G_M, G_{AF}, H_p)}{\Pr(G_C \mid G_M, G_{AF}, H_d)}.$$

Evaluation of the *PI* can proceed directly from this equation. The numerator can be evaluated using a Punnett square at each locus where both parents are present in the conditioning.

Assuming that the mother is the true mother it is often possible to determine the maternal and paternal alleles; A_m, and A_p unambiguously. This allows us to write:

$$\begin{aligned}\Pr(G_C \mid G_M, G_{AF}, H_d) &= \Pr(A_p, A_m \mid G_M, G_{AF}, H_d) \\ &= \Pr(A_m \mid G_M, G_{AF}, A_p, H_d)\Pr(A_p \mid G_M, G_{AF}, H_d)\end{aligned}.$$

Conventionally using the further assumption that

$\Pr(A_m \mid G_M, G_{AF}, A_p, H_d) = \Pr(A_m \mid G_M)$ allows the probability in the denominator of the *PI* to be written as

$$\Pr(G_C \mid G_M, G_{AF}, H_d) = \Pr(A_m \mid G_M)\Pr(A_p \mid G_M, G_{AF}, H_d)$$

Now $\Pr(A_m \mid G_M)$ is ½ or 1 depending on whether the genotype G_M containing the maternal allele is heterozygous or homozygous. We denote this probability as the maternal Mendelian factor M_M. Evaluation of $\Pr(A_p \mid G_M, G_{AF}, H_d)$ is slightly more problematic.

As with previous chapters we now turn to consideration of a series of examples and show in detail how to evaluate the paternity index, *PI*, for paternity trios.

Example 10.1

	Genotype
Mother	*cd*
Child	*ac*
Alleged father	*ab*

Under H_p we assume that the alleged father is the true father, and may proceed by using a Punnett square:

		Genes from the father	
		a	*b*
Genes from the mother	*c*	*ac*	*bc*
	d	*ad*	*bd*

We see that the child's genotype is one of the four (equiprobable) outcomes and assign the probability $\Pr(G_C \mid G_M, G_{AF}, H_p) = ¼$.

The mother is heterozygous for the maternal allele ($A_m = c$) and can assign the value $M_M = ½$ to the maternal Mendelian factor. The paternal allele is $A_p = a$. Under the hypothesis H_d we assign the probability $\Pr(A_p \mid G_M, G_{AF}, H_d) = p_a$, the allele probability of the *a* allele in this population. Hence the paternity index is

$$PI = \frac{\frac{1}{4}}{\frac{1}{2} \times p_a} = \frac{1}{2p_a}.$$

Example 10.2

	Genotype
Mother	*cc*
Child	*ac*
Alleged father	*ab*

Again under H_p we assume that the alleged father is the true father, and the Punnett square becomes:

		Genes from the father	
		a	*b*
Genes from the mother	*c*	*ac*	*bc*
	c	*ac*	*bc*

We see that the child's genotype occurs in two of the four (equiprobable) outcomes and assign the probability $\Pr(G_C \mid G_M, G_{AF}, H_p) = ½$.

The mother is homozygous for the maternal allele ($A_m = c$) and we can assign $M_M = 1$. The paternal allele $A_p = a$. As before we assign the probability $\Pr(A_p \mid G_M, G_{AF}, H_d) = p_a$ under the hypothesis H_d. Hence

$$PI = \frac{\frac{1}{2}}{1 \times p_a} = \frac{1}{2p_a}.$$

Example 10.3

	Genotype
Mother	*ab*
Child	*ab*
Alleged father	*bc*

Under H_p we assume that the alleged father is the true father, and may proceed by a Punnett square:

		Genes from the father	
		b	*c*
Genes from the mother	*a*	*ab*	*ac*
	b	*bb*	*bc*

We see that the child's genotype occurs in one of the four (equiprobable) outcomes and assign the probability ¼ to this genotype.

Table 10.3 Form of *PI* for all non-excluded combinations of maternal and paternal genotypes. Lee et al.[507]

Genotype Mother	Genotype Child	Genotype Alleged Father	*PI* (Alleged Father is True Father)
aa	*aa*	*aa*	$\frac{1}{p_a}$
ab			
bb	*ab*		
bc			
aa	*aa*	*ab*	$\frac{1}{2p_a}$
ab			
ac			
bb	*ab*		
bc			
bc	*ac*		
cc			
cd			
ab	*ab*	*aa*	$\frac{1}{p_a + p_b}$
		ab	
		ac	$\frac{1}{2(p_a + p_b)}$

This example was introduced because of a small complexity that occurs under H_d. This arises because either of the mother's alleles may be the maternal allele, making attribution of both the maternal and the paternal allele ambiguous. Under H_d we can see that the mother may contribute the *a* allele ($A_m = a$) with probability $M_M = ½$ or the *b* allele ($A_m = b$) with probability $M_M = ½$. If the maternal allele is $A_m = a$ then the paternal allele A_p must be *b*. If the maternal allele is $A_m = b$ then the paternal allele must be *a*. The denominator is therefore the sum of two terms. Hence

$$PI = \frac{\frac{1}{4}}{\frac{1}{2}p_a + \frac{1}{2}p_b} = \frac{1}{2(p_a + p_b)}.$$

There are 15 distinct combinations of maternal and paternal genotypes possible, but if we use the product rule to evaluate *PI* we find that *PI* takes only four possible forms, depending on whether the alleged father is a homozygote or a heterozygote and whether or not the child's paternal allele can be unambiguously identified.[507] In table 10.3 we tabulate the possible combination of mother, child and alleged father along with the *PI* formulae utilising the product rule.

An example from the 2003 exam

Q4. In a paternity dispute the mother, M, claims that a man AF is the father of the child, C.

	Locus 1	Locus 2
M	cd	ff
C	ac	ef
AF	ab	ee

Allele probabilities			
Locus 1		Locus 2	
a	0.10	e	0.22
b	0.08	f	0.20
c	0.12		
d	0.15		

a. Please give the mathematical definitions for the terms exclusion probability, the paternity index and the probability of paternity. Use the data from locus 1 to give examples of these terms. [5 marks]

PE is the probability that a random man would be excluded.

$PE_1 = (1-p_a)^2 = (1-0.1)^2 = 0.81$ $PE = 1-\prod_l (1-PE_l)$ (numerical result not requested but it does show understanding) part marks for $PE_1 = (1-\sum_i p_i)^2$ without the above

$PI_1 = \frac{\Pr(G_C, G_M, G_{AF} \mid Hp)}{\Pr(G_C, G_M, G_{AF} \mid Hp)} = \frac{1}{2p_a} = 5$ part marks for showing understanding by showing Punnet square or getting numerator or denominator correct (numerical result not requested but it does show understanding).

Prob pat $\frac{PI}{1+PI}$ =0.83 (numerical result not requested but it does show understanding).

1.5 marks off for each section incorrect. Most trouble was in PE. Part marks awarded for understanding.

b. Please give the formula for the paternity index at locus 2 in terms of the allele probabilities. [2 marks]

$PI_2 = \frac{\Pr(G_C, G_M, G_{AF} \mid Hp)}{\Pr(G_C, G_M, G_{AF} \mid Hp)} = \frac{1}{p_e}$ 1 mark for numerator correct 1 for denominator.

c. Please evaluate the paternity index at both loci. [1 marks]

$PI = \frac{1}{2 p_a p_e}$ or 5 and 4.55 =22.73 mark given for the algebraic solution or 5x4.55=22.73. Part marks off if the numbers are there but the multiplication is not done.

An example from the 2003 exam

Q2. In a paternity dispute the mother, M, claims that a man AF is the father of the child, C.

	Locus 1	Locus 2
M	cd	ff
C	ac	ef
AF	ab	ee

Allele probabilities			
Locus 1		Locus 2	
a	0.10	e	0.22
b	0.08	f	0.20
c	0.12		
d	0.15		

a. Please give the mathematical definitions for the terms exclusion probability, the paternity index and the probability of paternity. Use the data from locus 1 to give examples of these terms. [5 marks]

PE is the probability that a random man would be excluded.

$PE_1 = (1 - p_a)^2 \quad PE = 1 - \prod_l (1 - PE_l)$

$PI_1 = \frac{\Pr(G_C, G_M, G_{AF} \mid Hp)}{\Pr(G_C, G_M, G_{AF} \mid Hp)} = \frac{1}{2 p_a}$

Prob pat $\frac{PI}{1 + PI}$

b. Please give the formula for the paternity index at locus 2 in terms of the allele probabilities. [2 marks]

$PI_2 = \frac{\Pr(G_C, G_M, G_{AF} \mid Hp)}{\Pr(G_C, G_M, G_{AF} \mid Hp)} = \frac{1}{p_e}$

c. Please evaluate the paternity index at both loci. [1 marks]

$PI = \frac{1}{2 p_a p_e}$

An example from the 2004 exam

Please answer all parts.

a) The table below gives the genotypes of Tsar Nicholas II, Tsarina Alexandra, and a child.

vii) Please calculate the probability of exclusion, paternity index and probability of paternity for this child being a child of the Tsar and Tsarina. For the PI calculation use:

Hp: The child is a child of the Tsar and Tsarina

Hd: The child is a child of the Tsarina and a random man

	Loci	
	VWA	F13A1
Child	15,16	5,7
Tsar Nicholas II	15,16	7,7
Tsarina Alexandra	15,16	3,5
	Pr(15) = 0.10 Pr(16) = 0.15	Pr(3) = 0.05 Pr(5) = 0.06 Pr(7) = 0.07

	VWA	F13A1	Both
PE	0.5625	0.8649	0.940894
1-PE	0.4375	0.1351	0.059106

$$PI_{vWA} = \frac{\frac{1}{2}}{\frac{1}{2}\left(\Pr(15)+\Pr(16)\right)} = 4 \qquad PI_{F13A1} = \frac{\frac{1}{2}}{\frac{1}{2}\Pr(7)} = 14.3$$

$$PI_{both} = 4 \times 14.3 = 57.1$$

$$PP = \frac{PI}{1+PI} = \frac{57.1}{58.1} = 0.983$$

(10 marks)

viii) Please critique the value of these three methods of interpretation.
(2 marks)

PE wastes information. Possibly mention Brenner's example.

PP makes an assumption of equal prior odds which is often unjustified and potentially very wrong.

PI is the preferred method but can be difficult to explain.

c) Tsar Nicholas II had a brother Grand Duke Michael. What is the probability that Grand Duke Michael has the following genotype? Please show your working to justify your answer.

	Loci	
	VWA	F13A1
Tsar Nicholas II	15,16	7,7
Grand Duke Michael	15,16	3,5

(8 marks)

Not taught in 2009

An example from the 2006 Exam

6. Please answer all parts.

(a) In a paternity dispute the mother, M, claims that a man AF is the father of the child, C.

	Locus 1	Locus 2
M	ab	cd
C	aa	de
AF	ab	ef

Allele probabilities			
Locus 1		Locus 2	
a	0.10	c	0.12
b	0.08	d	0.15
		e	0.25
		f	0.20

(i) Please calculate the exclusion probability (PE), the paternity index (PI) and the probability of paternity (PP) for these two loci. Please include your workings as marks will be given for the correct method. (6 marks)

	Locus 1	Locus 2	Both
PE	0.81	0.5625	0.916875
1-PE	0.19	0.4375	0.083125

$$PI_{locus1} = \frac{\frac{1}{2}}{\frac{1}{2}\Pr(a)} = 10 \qquad PI_{locus2} = \frac{\frac{1}{2}}{\frac{1}{2}\Pr(e)} = 4$$

$$PI_{both} = 10 \times 4 = 40$$

$$PP = \frac{PI}{1+PI} = \frac{40}{41} = 0.976$$

(ii) The lawyer for the defendant (AF) suggests to you that the best method for you to use is the exclusion probability. How would you answer as in court?
(5 marks)

PE wastes information. Possibly mention Brenner's example.

(iii) The lawyer for the defendant suggests to you that the assumption of prior odds of 1:1 is inappropriate. The complainant had fallen asleep drugged at a party where there were 11 men. He asks you to rework the PP using prior odds of 1:10. What is your answer?
(2 marks)

Posterior odds = PI x prior odds

$posterior\ odds = 40 \times \frac{1}{10} = 4$ then change the odds to a probability $PP = \frac{4}{4+1} = 0.8$

(iv) In redirection the prosecutor states to you that you have given a PI of 10. He asks: "By this do you mean that it is 10 times more likely that the defendant is the father?" How would you answer? (2 marks)

This is an example of the prosecutors' fallacy. The statement is Pr(Hp|E). Points for stating this, and giving an explanation of PP is paternity and that it needs an assumption of prior odds and points for Ian's coping trick.

Example from the 2007 exam
Please answer all parts

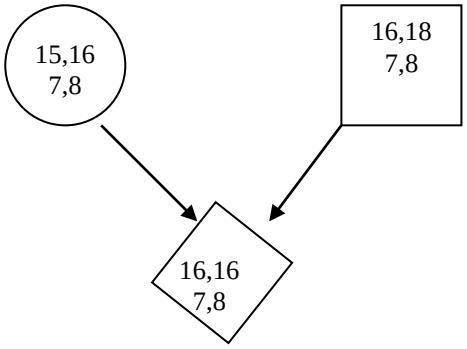
(a) What are Mendel's two laws? (2 marks)

(b) Please draw a small pedigree using the correct symbols, calculate the paternity index, probability of exclusion and probability of paternity for the following genotype data. Indicate where you use Mendel's laws. (8 marks)

Locus	Mother	Child	Alleged father
1	15,16	16,16	16,18
2	7,8	7,8	7,8

Allele probabilities locus 1	
15	0.10
16	0.12
18	0.15

Allele probabilities locus 2	
7	0.20
8	0.25



$$PI_{locus\,1} = \frac{\frac{1}{2}\times\frac{1}{2}}{\frac{1}{2}\times \Pr(16)} = \frac{1}{2\Pr(16)} = 4.17$$

$$PI_{locus\,2} = \frac{\frac{1}{2}}{\frac{1}{2}\times\left[\Pr(7)+\Pr(8)\right]} = \frac{1}{\Pr(7)+\Pr(8)} = 2.22$$

$$PI_{both} = 4.17\times 2.22 = 9.26$$

The factor's of ½ in the PI's come from Mendel's first law (segregation).
The multiplication when we do the PI for both loci comes from Mendel's 2nd law (independent assortment)

	Locus 1	Locus 2	Both loci
paternal allele(s)	16	7 ,8	
# alleles	1	2	
PE	$(1-\Pr(16))^2 = 0.7744$	$(1-\Pr(7)-\Pr(8))^2 = 0.3025$	0.8426
1-PE	0.2256	0.6975	0.1574

Probability of paternity $PP = \frac{PI}{1+PI} = 0.9025$

(c) Explain to a scientific audience the strengths and weaknesses of the probability of exclusion, probability of paternity, and paternity index. (5 marks)

PE wastes information. Possibly mention Brenner's example.

PP makes an assumption of equal prior odds which is often unjustified and potentially very wrong.

PI is the preferred method but can be difficult to explain.

Example from the 2009 exam

Q4i. A man is accused of fathering a child. The genotypes of the child, the mother and the alleged father are given below. Please evaluate the probability of exclusion, the paternity index and the probability of paternity for this case.
8 marks

	Locus 1	Locus 2
Mother	aa	cd
Child	ab	cd
Alleged father	bb	cc

You may use

allele	Pr(allele)
a	0.12
b	0.10
c	0.12
d	0.18

What are the drawbacks of the probability of exclusion? 2 marks

	Locus 1	Locus 2	
M	aa	cd	
C	ab	cd	
AF	bb	cc	
Paternal allele	b	c or d	
Number of paternal alleles n	1	2	
PE	0.81	0.49	0.9031
1-PE	0.19	0.51	0.0969

$$PI_{locus1} = \frac{1}{\Pr(b)} = 10 \; PI_{locus2} = \frac{1}{\Pr(c) + \Pr(d)} = 3.333 \; PI_{both} = 33.33$$

$$PP = \frac{33.33}{34.33} = 0.971$$

Drawbacks of the PE are that it wastes information. Specifically the genotype of the AF.